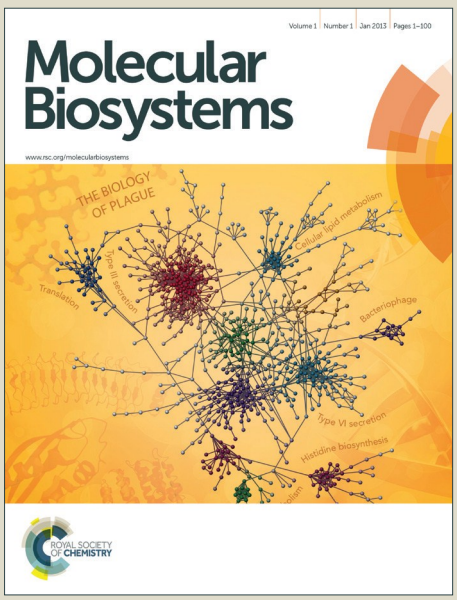
Molecular BioSystems

Accepted Manuscript



ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/molecularbiosystems

Submitted to Molecular BioSystems, 06-21-2015

Engineering Reprogrammable RNA-Binding Proteins for Study and Manipulation of the Transcriptome

Zhanar Abil[1] and Huimin Zhao[1,2]*

[1]Department of Biochemistry, [2] Department of Chemical and Biomolecular Engineering, Department of Bioengineering, Department of Chemistry, Center for Biophysics and Computational Biology and Institute for Genomic Biology

University of Illinois at Urbana-Champaign

600 South Mathews Avenue, Urbana, IL 61801, USA

KEYWORDS: RNA-binding Domain, RBD, RNA-binding protein, RBP, Pumilio, Pumilio and fem3 mRNA-binding factor, PUF, pentatricopeptide repeat, PPR, CRISPR/Cas

* To whom correspondence should be addressed.
E-mail: zhao5@illinois.edu,
Tel: +1-(217) 333-2631
Fax: +1-(217) 333-5052

Abstract

With the expanding interest in RNA biology, interest in artificial RNA-binding proteins (RBPs) is likewise increasing. RBPs can be designed in a modular fashion, whereby effector and RNA-binding domains are combined in chimeric proteins that exhibit both functions and can be applied for regulation of a broad range of biological processes. The elucidation of the RNA recognition code for Pumilio and fem-3 mRNA-binding factor (PUF) homology proteins allowed engineering of artificial RBPs for targeting endogenous mRNAs. In this review, we will focus on the recent advances in elucidating and reprogramming PUF domain specificity, update on several promising applications of PUF-based designer RBPs, and discuss some other domains that hold the potential to be used as the RNA-binding scaffolds for designer RBP engineering.

Introduction

RNA-protein interactions are ubiquitous in biology, and play a key role in numerous processes required for regulation of gene expression. In eukaryotes, this regulation is achieved by controlling multiple aspects of mRNA metabolism such as biogenesis, transport, maturation, and turnover[1, 2]. In addition, mRNAs are in many cases co-regulated by noncoding RNA, through collaborative or competitive interplay between noncoding RNA and regulatory proteins[3]. Finally, the translation apparatus itself requires post-transcriptional modification of RNA bases in rRNAs and tRNAs, which involves RNA-protein interactions[4]. In all of these phenomena, recognition and binding of RNA by RNA-binding proteins (RBPs) is imperative. However, we have only scratched the surface in elucidating these processes, and hence, our understanding of RNA biology and ability to manipulate it are currently very limited. In the fashion of synthetic biology, it is proposed that engineering of designer RBPs can not only enable us with useful new technologies for precise gene regulation, but also deepen our understanding of RNA biology.

Most RBPs have a modular configuration, where RNA-binding domains (RBDs), protein-binding domains, and catalytic domains are usually physically separate and function independently from one another[5, 6]. Emulating Nature, artificial RBPs can be created in a similar fashion, by combining effector domains of choice with suitable RBDs[6-10]. These designer RBPs can be implemented in the regulation of a broad range of biological processes that include RNA metabolism, including regulation of mRNA translation, degradation, splicing, editing, and localization.

For the targeting of untagged endogenous RNA, RBDs with easily reprogrammable specificity would be required. Such remarkable trait was found in Pumilio and fem-3 mRNA-binding factor (PUF) homology proteins[11, 12], as well as pentatricopeptide repeat proteins (PPR)[13-15]. In the past decade, a lot has been elucidated about the RNA-binding nature of PUF proteins, which allowed implementation of this domain for the engineering of PUF-based designer RBPs[9, 10]. In this brief review, we will focus on the recent advances in elucidating and reprogramming PUF domain specificity, update on several promising applications of PUF-based designer RBPs, and briefly discuss other systems (PPR proteins and the CRISPR/Cas9 system) that hold the potential to be used as RNA-binding scaffolds for designer RBP engineering.

PUF Domain: Reprogrammable RNA-Binding Protein

PUF family proteins are named after the founding members Pumilio in *Drosophila melanogaster*[16] (DmPum) and fem-3 mRNA-binding factor (FBF) in *Caenorhabditis elegans*[17]. Since their initial discovery, PUF homology proteins have been found in all eukaryotic organisms[18, 19]. PUF domains usually bind to the 3′ untranslated region (UTR) of their target transcripts, and regulate mRNA stability, translation, cellular localization (reviewed elsewhere[19, 20]), pre-rRNA processing[21], and recognition of viral infections[22, 23].

PUF family proteins are typically composed of 8 imperfectly repeated 36 amino acid motifs (PUF repeats), which, together with flanking conserved sequences, form a sequence-specific single-stranded (ss) RNA-binding domain, the

PUF domain[24, 25]. The conserved sequences flanking PUF repeats (Csp1 and Csp2 motifs, which typically precede and follow the PUF repeats, respectively) are composed of degenerate repeats and are required for RNA binding in DmPum and *C. elegans* FBF.[24, 25] The crystal structures of PUF domains revealed that the eight PUF repeats are very similar structurally, and each of them is comprised of a triangle-shaped structural unit consisting of three helical segments[26, 27]. These repeated units typically pack together to form a crescent-shaped right-handed superhelix with a continuous hydrophobic core[26, 27].

The crystal structure of a PUF domain from human Pumilio 1 (HsPum1) protein in a complex with RNA revealed its modular RNA-recognition mode[11]. RNA is bound to the concave surface of the protein, where each protein repeat interacts with a single RNA base. The N-terminal repeat of the PUF domain interacts with the 3′ end of the mRNA sequence, thus the PUF domain binds RNA in an "antiparallel" configuration[11] (Figure 1a). Repeats 2, 6, and 8 interact with three uracil bases present in the target RNA, where an asparagine residue at position 12 and a glutamine residue at position 16 make hydrogen bonds with the O2, N3, and O4 groups of the Watson-Crick face of the bases (Figure 2a). Repeats 1, 3, and 5, on the other hand, interact with adenines, where a cysteine or a serine at position 12 forms a van der Waals interaction of the C2 group of the base, and a glutamine at position 16 makes hydrogen bonds with the N1 and N6 groups of the Watson-Crick face of the bases (Figure 2a).

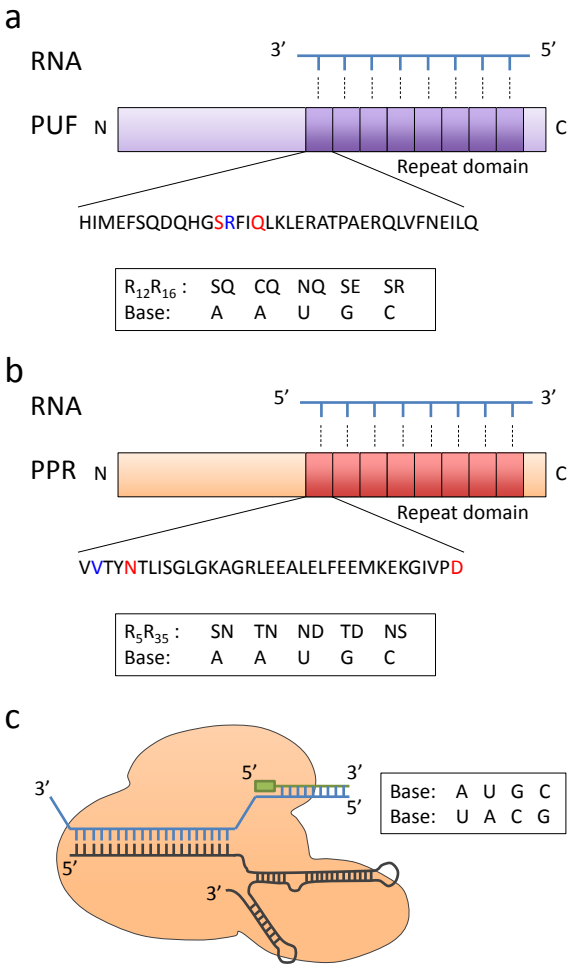


Figure 1. Schematics of RNA-binding proteins **(a)** Schematic of PUF binding to ssRNA. The primary sequence of the first repeat of HsPum1 is shown. Recognition code is shown in a box. Red, amino acids at positions 12 and 16 that interact with the Watson-Crick edge of RNA bases; blue, stacking amino acid residue at position 13. **(b)** Schematic of PPR binding to ssRNA. An engineered repeat constructed from evolutionarily conserved amino acids (the schematic is adapted from the reference 28) is shown. Note that the numbering system used in the figure is from the reference 14. Alternative numbering systems include that used in references 13, 29, where the marked residues 5 and 35 correspond to residues 6 and 1′, and that proposed in the reference 15, where they correspond to residues 4 and ii, respectively. Red, amino acids at positions 5 and 35 that contribute to recognition of RNA bases; blue, stacking amino acid residue at position 2. **(c)** Schematic of RNA-guided Cas9 recognition of ssRNA. Blue, target RNA sequence; green, PAM-presenting DNA oligonucleotide; Grey, guide RNA sequence (adapted from reference 30).

Repeat 7 interacts with the only guanine in the structure, where a serine at position 12 and a glutamic acid residue at position 16 make hydrogen bonds with the Watson-Crick face of the base. A tyrosine or an arginine at position 13 generally form stacking interactions between aromatic rings of adjacent RNA bases[11]. An altered positioning of the fifth base with respect to the glutamine residue at position 16 in repeat 4 allows repeat 4 to interact with the Watson-Crick edges of both uracil and cytosine[11], as well as form van Waals interactions with the Hoogsteen edges of adenine and guanine[31].

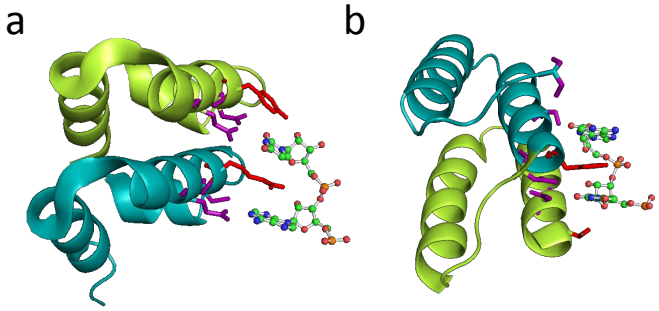


Figure 2. Crystal structures of PUF and PPR motifs. **(a)** Structure of PUF repeats 2 (lemon) and 3 (teal) from HsPum1 (PDB 1MW8) recognizing uracil and adenine, respectively. Amino acid residues at positions 12 and 16, which confer RNA base specificity are shown in purple. Residues at position 13, which form stacking interactions between bases are shown in red. **(b)** Structure of PPR motifs 4 (lemon) and 5 (teal) from maize PPR10 (PDB 4M59) recognizing uracil and adenine, respectively. Amino acid residues at positions 5 and 35 that confer RNA base specificity are shown in purple. Residues at position 2 that form stacking interactions between bases are shown in red.

A general RNA-recognition code was proposed based on this co-crystal structure, where the $R_{12}R_{16}$ combinations NQ, CQ/SQ, and SE recognize uracil, adenine, and guanine, respectively. No code for cytosine was apparent from this structure. The proposed RNA-recognition code suggested that PUF domains with

designed specificities could be created via site-directed mutagenesis. Indeed, mutagenesis of only 2 or 3 (including the "stacking" amino acid) residues in a repeat was sufficient to predictably alter PUF domain specificity in HsPum, as confirmed by *in vitro* binding assays[11, 12, 32].

Identification of the cytosine-recognition code of PUF repeats expanded the RNA-binding specificity of designer PUF domains to recognize any RNA target sequence[33, 34]. In order to select for a PUF repeat with specificity for cytosine, the interaction between the PUF domain and target RNA was linked to a life-death selection in *Saccharomyces cerevisiae* using a yeast three-hybrid system. From a pool of PUF domain variants with randomized amino acids at positions 12 and 16 in repeat 6, PUF repeats with the residue combination SR were selected[33]. In a similar study, arginine at position 16 was likewise selected; while in position 12, other small or nucleophilic side chains (G, A, S, T, or C) were selected[34]. This code could be transferred to other PUF repeats[33, 34], and PUF domains with multiple cytosine-recognizing repeats were shown to be functional[33].

Recently, the PUF specificity code was expanded further using data gathered from a high-throughput sequencing approach. These data allowed determination of the specificity conferred by various combinations of the three amino acid residues introduced into the seventh repeat of FBF-2[35]. This approach, termed SEQRS, combined *in vitro* selection, high-throughput sequencing of RNA and sequence specificity landscapes. Multiple highly specific combinations were discovered for the recognition of uracil, adenine, and guanine, some of which were *de novo* designed amino acid combinations. For instance, the *de novo* designed CFQ and CYE

($R_{12}R_{13}R_{16}$) amino acid combinations were more specific for adenine than any other natural combinations[35].

In order to facilitate the mutagenesis of PUF domains, we developed a PUF repeat library with four variations in each PUF repeat, using the HsPum1 PUF domain as a scaffold[32]. Inspired by the Golden Gate assembly strategy commonly used for the assembly of transcription activator-like (TAL) effector proteins[36-39], we have similarly assembled designer PUF domains from up to 8 mutant repeats, and confirmed the change of specificity to the new predicted target RNA *in vitro*. The assembly of the gene is highly efficient and requires only 3 days for the generation of a PUF mutant of choice with as many mutations as necessary in accordance with the PUF RNA-recognition code[32].

Challenges of PUF Engineering

Despite the apparent simplicity of the described RNA-binding code of HsPum1, many PUF proteins utilize alternate binding modes, which may complicate the engineering of synthetic RBPs. For example, while crystal structures of PUF domains from human, mouse, fly, worm, and yeast PUF proteins revealed eight PUF repeats[11, 26, 27, 40-43], some PUF domains, such as human Puf-A, have an eleven-repeat fold shaped in an L-like structure[21]. Interestingly, Puf-A and its yeast ortholog Puf6 bind double or single-stranded RNA or DNA without sequence specificity in a mode different from classical PUF proteins[21]. In these atypical PUF domains, the protein/RNA interaction is achieved not through the C-terminal repeats that

resemble those of the eight-repeat PUF proteins, but through the basic residues on the inner concave surface of the N-terminal extension of Puf-A/Puf6[21].

More commonly, a number of PUF proteins are known to bind longer RNA sequences and exhibit relaxed specificity due to base flipping[40, 42, 44-47] and the use of additional RNA-binding pockets[43, 48]. Additional binding modes were described in HsPum1 and HsPum2, where recognition of the Hoogsteen edge of bases or base omission resulted in acceptance of bases not predicted by the code[31]. Therefore, it is desirable to engineer a PUF architecture that would minimize additional binding modes and allow the most predictable, robust scheme of engineering RNA specificity. To this end, development of standardized PUF repeats would facilitate applicability of PUF proteins for designer RBP engineering.

Furthermore, applications that would require targeting of a single RNA species in an entire transcriptome would necessitate higher PUF specificity. Wild type human Pumilio proteins HsPum1 and HsPum2 are known to have hundreds of RNA targets[49, 50]. The RBD architecture composed of 8 repeats that recognizes an RNA sequence of 8 nucleotides (nt) (which would recognize a sequence approximately once in 65,000 nt), appears to be suitable for this purpose. However, targeting a unique RNA sequence would require recognition of a target at least 14-16 nt-long (depending on the estimated size of the transcriptome), and further engineering of PUF to meet this requirement.

The challenge of engineering a repeat domain with higher nucleic acid specificity lies in the fact that each repeat has a characteristic binding energy. The cumulative binding energy of 8 PUF repeats might already be enough for binding to

the cognate 8 nt sequence and activity. Since additional PUF repeats would not necessarily interfere with the binding of the original 8 repeats, increasing the number of PUF repeats in a single protein may not necessarily lead to increased specificity. On the contrary, the specificity might decrease due to off-target binding of the additional repeats. In fact, it was shown with TAL effector DNA-binding repeat proteins that excess DNA-binding affinity due to a greater number of repeats leads to increased TAL effector nuclease activity against off-target sites[51].

Nevertheless, in the attempt of increasing PUF specificity, a PUF domain composed of 16 repeats was indeed developed. To this end, the HsPum1 RBD was inserted between repeats 5 and 6 of the same protein[34]. Using a yeast three-hybrid assay, it was demonstrated that the extended PUF domain bound to its cognate extended RNA target better than the eight-repeat PUF to the same RNA sequence[34]. Although demonstration of functionality of such an extended PUF domain is impressive, the increased specificity and reduced off-target effects of such architecture are yet to be systematically tested.

However, it is possible, in theory, to increase the PUF specificity by the implementation of a split-protein strategy, as shown previously[52-54]. In this approach, the functional protein is assembled from inactive parts only when two PUF domains fused to the inactive parts are bound to two adjacent binding sites, thus theoretically increasing the PUF specificity to 16 nt. However, not every protein has been shown to be functional when re-assembled in this fashion, so such a strategy might require idiosyncratic engineering of a number of split protein systems. It is therefore desirable to develop a PUF architecture with reduced

binding energy per repeat, which will only bind to an RNA target when specifically bound to 16 nt.

PUF Domain in Applications

The first application of a PUF-based designer RBP was in live-cell tracking of mRNA. Visualization of RNA in live cells has previously been possible by tagging the RNA of interest with multiple copies of recognition elements and the use of RBPs tagged with green fluorescent protein (GFP) or similar proteins[55]. Although successful in detecting and tracking RNA in live ells, this system requires prior tagging of the RNA of interest, which may alter RNA abundance, translation, and localization. Application of reprogrammable RBPs such as PUF domains for this purpose allows targeting of untagged (and therefore minimally perturbed) endogenous RNA. In order to reduce background fluorescence and improve the signal-to-noise ratio, split-protein systems were implemented. For that purpose, two PUF domains were used to reconstitute enhanced GFP (EGFP) or Venus from two nonfunctional parts that were brought to proximity on target RNA[52] (Figure 3a). This approach allowed tracking of mitochondrial RNA[52] or single molecules of β-actin mRNA[54] in cultured mammalian cells. In a similar strategy, a split mCitrine-PUF system was used to track tobacco mosaic virus RNA in plant cells[53].

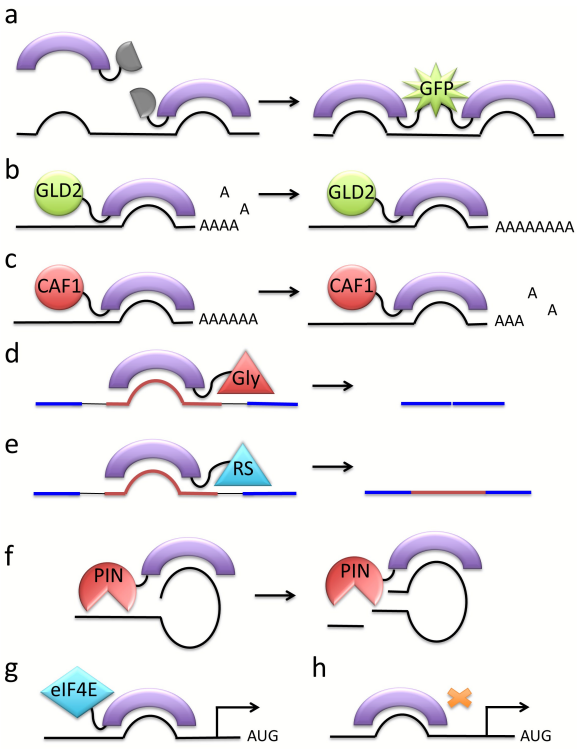


Figure 3. Schematics of reported PUF applications. **(a)** RNA visualization using a split GFP system. **(b)** Polyadenylation of mRNA using fusion to GLD2. **(c)** Deadenylation of mRNA using fusion to CAF1. **(d)** Intron exclusion using a Gly-rich domain. **(e)** Intron inclusion using an Arg-Ser-rich domain. **(f)** RNA cleavage using the PIN domain. **(g)** Enhancement of translation using eIF4E. **(h)** Translation inhibition using PUF as a steric block to the translation initiation machinery.

The ability to control gene expression is necessary for both fundamental research as well as in applications ranging from cell fate manipulation to metabolic engineering. Gene expression is regulated via the control of multiple aspects of RNA metabolism, including capping, polyadenylation, splicing, translation, degradation, editing, and transport. By fusing corresponding functional domains to RBDs such as PUF, one can envision development of tools capable of sequence-specific manipulation of RNA almost at any processing step (Figure 3b-h).

One such tool has been developed to inhibit translation via regulation of the poly(A) tail of mRNA. To inhibit translation of microinjected reporter mRNA in *Xenopus* oocytes, FBF-2, a PUF domain found in *Caenorhabditis elegans,* was fused to *Xenopus* CAF1 protein, which is known to exhibit deadenylation and translation inhibition activities[56]. The system induced deadenylation and decreased translation of luciferase mRNA. In another approach, a PUF domain alone, targeted to the 5′ UTR of an mRNA, was used as a steric block in the translation initiation pathway to inhibit the translation of a downstream reporter gene[57, 58]. The elegance of such approach allowed for translational repression of an individual reporter cistron in a bicistronic reporter in *Escherichia coli*[58]. In our work, a PUF domain from HsPum1 was fused to human tristetraprolin (TTP), a protein involved in AU-rich element-mediated translational repression and mRNA degradation. The system allowed not only translational inhibition of reporter mRNA, but also translational inhibition of endogenously expressed vascular endothelial growth factor A (VEGFA) in cultured human cells[32].

The use of a PUF scaffold for sequence-specific enhancement of gene expression in living systems was first demonstrated in *Xenopus* oocytes. FBF-2 protein was fused with GLD2, a cytoplasmic poly(A) polymerase, and the chimeric protein was shown to direct polyadenylation of reporter RNA, enhance translation of microinjected luciferase mRNA, and direct polyadenylation of endogenous ribosomal protein L1 mRNA[56]. In another well-designed study, the PUF domain, which was targeted to the 5′ UTR of mRNA, was fused to the eukaryotic translation initiation factor 4E (eIF4E) to enhance translation of reporter luciferase mRNA[57].

Furthermore, light-inducible heterodimerization of PUF and eIF4E through light-sensitive protein partners was demonstrated, thus allowing light-inducible translation activation[57]. Alternatively, the PUF domain of FBF-2 protein that was mutagenized for specific binding to a new target sequence was fused with a segment of *S. cerevisiae* poly(A)-binding protein, which is known to stimulate translation. This chimeric protein, referred to as "neo-activator", increased endogenous cyclin B1 abundance in a cancer cell line, and rendered the cell line hypersensitive to chemotherapeutic drugs[35].

Nature's other strategy for controlling gene expression is regulation of mRNA abundance. In many organisms, this strategy is achieved through the RNA interference (RNAi) pathway, where a multi-component ribonucleoprotein complex sequence-specifically cuts and degrades the target RNA[59]. An alternative synthetic system for RNA cleavage termed artificial site-specific RNA endonuclease (ASRE) was designed using a PUF domain fused to a non-specific endonuclease domain PilT N-terminus (PIN)[60]. As a demonstration of applicability of ASRE in systems where the RNAi pathway is absent, ASREs were designed to recognize and cut the endogenous *LacZ* transcript in *E. coli* and a mammalian mitochondrial gene[60]. ASRE was further developed for the potential treatment of myotonic dystrophy type 1 (DM1)[61]. Caused by the expansion of $(CTG)_n$ in the 3′ UTR of dystrophia myotonica-protein kinase (*DMPK*) gene, DM1 is characterized by the accumulation of $(CUG)_n$-containing transcripts in the nucleus. ASRE was designed to specifically bind and cleave $(CUG)_n$-repeat containing transcripts in DM1 patient cells and reverse the phenotypical abnormalities associated with the pathogenic *DMPK* mRNA[61].

Alternative splicing is an important step in eukaryotic mRNA processing that allows for expression of multiple isoforms of proteins with distinct functions. The PUF domain was used as a scaffold for engineering artificial splicing factors with designed sequence specificities and activities[33, 62]. These engineered splicing factors (ESFs) were constructed from a wild type or modified PUF domain of HsPum1 fused to a glycine-rich domain of hnRNP A1 (Gly-PUF) or the arginine-serine-rich domain of ASF/SF2 (RS-PUF). Consistent with the reported activities of the tethered domains, a Gly-PUF type ESF repressed inclusion of the cassette exon containing the cognate target sequence of PUF, while an RS-PUF type ESF promoted inclusion of the same cassette exon in reporter constructs[33, 62]. The designer ESF was shown to be able to shift splicing of endogenous pre-mRNA of Bcl-x towards the short Bcl-xS isoform, thereby promoting apoptosis[62], as well as promote splicing towards the anti-angiogenic isoform b of endogenous VEGFA gene[33] in cultured cancer cells.

The plug and play approach of PUF-based RBP engineering is still far from being exhausted. For example, in our unpublished work, we engineered a ligand-dependent interaction of the PUF domain with the cellular motor kinesin. Using this system, we observed re-localization of reporter mRNA containing PUF-binding sites in its 3′ UTR to the (+)-end of microtubules anchored at the cell periphery of mammalian cells (Abil et al., unpublished results). This tool could be useful in regulation of local mRNA translation in highly polarized cells like neurons. Other potential applications of PUF-based factors have been proposed[7, 9], and have not yet been demonstrated. These include factors to control RNA localization, RNA editing,

control of alternative polyadenylation, regulation of transcription, and alternative approaches or improved versions for already demonstrated applications.

Potential Use of Other Reprogrammable RNA-Binding Proteins

PPR proteins are another RNA-binding class of proteins found primarily in mitochondria and chloroplasts of land plants, where they participate in RNA cleavage, splicing, degradation, editing and translation[63, 64]. The RNA-binding motif consists of 2-30 degenerate repeats that are approximately 35 amino acids long[65] and are organized in a helix-turn-helix structure[14, 66, 67]. Like PUF proteins, PPR proteins bind RNA in a 1:1 correspondence between repeats and bases, and recognize ssRNA in a modular fashion[13-15, 68]. Computational, biochemical, and structural analyses of PPR proteins agree that residue 5 in each repeat contributes to distinguishing between purines and pyrimidines[13-15, 68]. Residue 35, on the other hand, was predicted to discriminate between adenine and guanine, as well as between cytosine and uracil. It should be noted that alternative numbering of amino acid residues in the PPR motif has been used by different authors. The preferred numbering used in this review is that used in the reference 14, where the residue delimitation started from the first residue in the first helix of a repeat. Alternative numbering systems include that used in references 13 and 29, where the aforementioned residues 5 and 35 of the same repeat correspond to a residue 6 of the same repeat and 1' of the following repeat, and that proposed in the reference 15, where they correspond to residues 4 and ii (2 residues before the first amino acid of the next repeat), respectively.

The co-crystal structure of maize chloroplast PPR10 with an 18-nt RNA from *PSAJ-RPL33* intergenic region provided a molecular basis for specific and modular recognition of adenine, guanine, and uracil[14]. Specific, modular recognition of *PSAJ* RNA is achieved by repeats 3-6 and repeats 16 and 17 of the 19-repeat PPR domain. In each of these repeats, residue 5 interacts with a single RNA base (5'-G1-U1-A3-U4-3' and 5'-U15-U16-3') through direct hydrogen bonds. Thus, residues threonine, asparagine, and serine at position 5 in these repeats are bound to guanine, uracil, and adenine[14] (Figure 2b). Another amino acid, which is located at position 2, is sandwiched between two adjacent bases in a stacking interaction[14, 29], and thus is analogous to the residue at position 13 in PUF proteins (Figure 2b). Residue 35 is located in the vicinity of the recognized base (Figure 2b), and is speculated to contribute to binding either through water molecules or by stabilizing the residue 35 at an RNA-bound conformation[14]. Noteworthy, the specific recognition of RNA by a PPR motif is "parallel", i.e. the N-terminal repeat of PPR binds to the 5′-end of the target RNA sequence, in contrast to the PUF proteins (Figure 1b). In addition, PPR10 in complex with RNA forms an asymmetric dimer, the existence/significance of which under physiological conditions remains unknown[14].

The structure of a small PPR from *Brachypodium distachyon*, thylakoid assembly 8 (THA8), revealed five tandem repeats[69]. Only one of these repeats, repeat 4, exhibited a specific interaction with a base as predicted by the code: guanine recognition by residues threonine and aspartate at positions 5 and 35, respectively[69]. The rest of the RNA is nonspecifically bound to the asymmetric dimer interface formed by two THA8 monomers. Such overall non-modular interaction of

PPR with RNA is speculated to have arisen due to an inherently small RNA recognition capacity of short PPRs, dimerization of which could have extended the RNA-binding surface and increased the affinity towards RNA[69]. Overall, these results suggest that short PPRs (5 repeats and shorter) may not exhibit sufficient RNA recognition capacity for biotechnological applications.

In vitro binding assays confirmed that mutagenesis of these two amino acids could change the RNA sequence specificity according to the code[13]. The RNA-recognition code was further confirmed using designer PPR proteins constructed from sequences of standardized artificial repeats[28, 70]. The standardized repeats were constructed by selection of the most evolutionarily conserved amino acids at each position of the PPR. Eight[28] or ten[70] of these standardized repeats were assembled to form a designer PPR domain, which, unlike natural PPR proteins, was shown to be highly soluble[28]. Using an *in vitro* binding assay, the code for specific recognition of adenine, uracil, and cytosine was confirmed as R_5R_{35} combinations ND for uracil, TN or SN for adenine, NS for cytosine[28, 70]. This artificial PPR architecture allowed specific and modular recognition of RNA *in vitro*[28, 70] and is promising for the development of designer RBPs for use in live cells.

Due to a more recent elucidation of the RNA-recognition code, PPR proteins have not yet been used as widely as PUF proteins in reprogramming RNA specificity. However, unlike PUF proteins, PPR proteins differ widely in their recognized RNA sequences, which may be an indication of a greater plasticity and amenability to engineering specificity.

Clustered regularly interspaced short palindromic repeats (CRISPR)/CRISPR-associated (Cas) systems are part of the bacterial and archaeal adaptive immune system that targets phage DNA[71-74]. In this system, Cas9 binds and cleaves foreign DNA targets following the hybridization of the CRISPR-RNA to DNA. In addition, Cas9 recognition and cleavage activities require a short DNA sequence known as the protospacer adjacent motif (PAM), which is located next to and on the opposite strand of the 20-nt target sequence[75]. Because of this ability to bind and cleave double-stranded DNA, as well as its simple re-programmability, the CRISPR/Cas system has been used as a powerful tool for genome editing and gene expression regulation in multiple prokaryotic and eukaryotic systems[76].

Recently, it was demonstrated that the CRISPR/Cas system can also be used for specific binding and cleavage of RNA *in vitro*[30]. For this purpose, *Streptococcus pyogenes* Cas9 was used in combination with guide RNA and a separate ssDNA oligonucleotide (PAMmer) bearing the PAM motif (Figure 1c). The specially designed PAMmer enabled binding and cleavage of target ssRNA while avoiding the corresponding DNA sequences. In addition, a nuclease-deficient Cas9 could be used to pull down endogenous, untagged GAPDH mRNA from HeLa cell lysates. The applicability of the CRISPR/Cas system for RNA recognition and functional activity has not yet been demonstrated *in vivo*. However, this strategy of reprogramming RNA recognition is promising, and if successful, will abolish the need of redesigning RBPs for each new target.

Conclusions

Our interest in RNA biology is increasing, and so is the demand for new tools for the study and manipulation of RNA. Despite the relative youth of the reprogrammable RBD field, exciting advances have been made in the past decade. PUF is no longer the only domain for which engineering of specificity has been demonstrated. The discovery of the recognition code of the PPR domain as well as implementation of Cas9 for RNA recognition made them promising candidates for engineering designer RBPs. To date, however, only PUF domains have been extensively demonstrated as versatile tools that can function in live cells. The next several years will reveal if PUF domains will still be the RBD of choice, or if novel technologies will take over. In either case, the designer RBD field is clearly gaining momentum, which will lead to the diversification of the field and powerful new tools for regulation of the transcriptome.

Author Information

Corresponding Author

*(H.Z.) E-mail: zhao5@illinois.edu. Tel: +1-(217) 333-2631. Fax: +1-(217) 333-5052.

Author Contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

Acknowledgements

We gratefully acknowledge financial support from the Energy Biosciences Institute, National Institutes of Health (GM077596), Defense Advanced Research Projects

Agency (DARPA), Department of Energy (DE-SC0008743), and National Academies Keck Futures Initiative on Synthetic Biology.

References

1. S. Kishore, S. Luber and M. Zavolan, *Brief Funct Genomics*, 2010, 9, 391-404.
2. A. Re, T. Joshi, E. Kulberkyte, Q. Morris and C. T. Workman, *Methods Mol. Biol.*, 2014, 1097, 491-521.
3. J. J. Ho and P. A. Marsden, *Wiley Interdiscip Rev RNA*, 2014, 5, 69-86.
4. R. Ishitani, S. Yokoyama and O. Nureki, *Curr. Opin. Struct. Biol.*, 2008, 18, 330-339.
5. B. M. Lunde, C. Moore and G. Varani, *Nat. Rev. Mol. Cell Biol.*, 2007, 8, 479-490.
6. A. Filipovska and O. Rackham, *RNA Biol.*, 2011, 8, 978-983.
7. J. P. Mackay, J. Font and D. J. Segal, *Nat. Struct. Mol. Biol.*, 2011, 18, 256-261.
8. Y. Chen and G. Varani, *FEBS J*, 2013, 280, 3734-3754.
9. Y. Wang, Z. Wang and T. M. Tanaka Hall, *FEBS J*, 2013, 280, 3755-3767.
10. R. Choudhury and Z. Wang, *Adv. Exp. Med. Biol.*, 2014, 825, 199-225.
11. X. Wang, J. McLachlan, P. D. Zamore and T. M. Hall, *Cell*, 2002, 110, 501-512.
12. C. G. Cheong and T. M. Hall, *Proc. Natl. Acad. Sci. U. S. A.*, 2006, 103, 13635-13639.
13. A. Barkan, M. Rojas, S. Fujii, A. Yap, Y. S. Chong, C. S. Bond and I. Small, *PLoS Genet*, 2012, 8, e1002910.
14. P. Yin, Q. Li, C. Yan, Y. Liu, J. Liu, F. Yu, Z. Wang, J. Long, J. He, H. W. Wang, J. Wang, J. K. Zhu, Y. Shi and N. Yan, *Nature*, 2013, 504, 168-171.
15. Y. Yagi, S. Hayashi, K. Kobayashi, T. Hirayama and T. Nakamura, *PLoS One*, 2013, 8, e57286.
16. R. Lehmann and C. Nusslein-Volhard, *Development*, 1991, 112, 679-691.
17. J. Ahringer and J. Kimble, *Nature*, 1991, 349, 346-348.
18. M. Wickens, D. S. Bernstein, J. Kimble and R. Parker, *Trends Genet.*, 2002, 18, 150-157.
19. T. Quenault, T. Lithgow and A. Traven, *Trends Cell Biol.*, 2011, 21, 104-112.
20. M. A. Miller and W. M. Olivas, *Wiley Interdiscip Rev RNA*, 2011, 2, 471-492.
21. C. Qiu, K. L. McCann, R. N. Wine, S. J. Baserga and T. M. Hall, *Proc. Natl. Acad. Sci. U. S. A.*, 2014, 111, 18554-18559.
22. S. Un Huh and K. H. Paek, *Plant Signal Behav*, 2013, 8, e23975.
23. R. Narita, K. Takahasi, E. Murakami, E. Hirano, S. P. Yamamoto, M. Yoneyama, H. Kato and T. Fujita, *PLoS Pathog.*, 2014, 10, e1004417.
24. P. D. Zamore, J. R. Williamson and R. Lehmann, *RNA*, 1997, 3, 1421-1433.
25. B. Zhang, M. Gallegos, A. Puoti, E. Durkin, S. Fields, J. Kimble and M. P. Wickens, *Nature*, 1997, 390, 477-484.
26. T. A. Edwards, S. E. Pyle, R. P. Wharton and A. K. Aggarwal, *Cell*, 2001, 105, 281-289.
27. X. Wang, P. D. Zamore and T. M. Hall, *Mol. Cell*, 2001, 7, 855-865.

28. S. Coquille, A. Filipovska, T. Chia, L. Rajappa, J. P. Lingford, M. F. Razif, S. Thore and O. Rackham, *Nat Commun*, 2014, 5, 5729.
29. S. Fujii, C. S. Bond and I. D. Small, *Proc. Natl. Acad. Sci. U. S. A.*, 2011, 108, 1723-1728.
30. M. R. O'Connell, B. L. Oakes, S. H. Sternberg, A. East-Seletsky, M. Kaplan and J. A. Doudna, *Nature*, 2014, 516, 263-266.
31. G. Lu and T. M. Hall, *Structure*, 2011, 19, 361-367.
32. Z. Abil, C. A. Denard and H. Zhao, *J. Biol. Eng.*, 2014, 8, 7.
33. S. Dong, Y. Wang, C. Cassidy-Amstutz, G. Lu, R. Bigler, M. R. Jezyk, C. Li, T. M. Hall and Z. Wang, *J. Biol. Chem.*, 2011, 286, 26732-26742.
34. A. Filipovska, M. F. Razif, K. K. Nygard and O. Rackham, *Nat. Chem. Biol.*, 2011, 7, 425-427.
35. Z. T. Campbell, C. T. Valley and M. Wickens, *Nat. Struct. Mol. Biol.*, 2014, 21, 732-738.
36. T. Cermak, E. L. Doyle, M. Christian, L. Wang, Y. Zhang, C. Schmidt, J. A. Baller, N. V. Somia, A. J. Bogdanove and D. F. Voytas, *Nucleic Acids Res.*, 2011, 39, e82.
37. T. Li, S. Huang, X. Zhao, D. A. Wright, S. Carpenter, M. H. Spalding, D. P. Weeks and B. Yang, *Nucleic Acids Res.*, 2011, 39, 6315-6325.
38. R. Morbitzer, J. Elsaesser, J. Hausner and T. Lahaye, *Nucleic Acids Res.*, 2011, 39, 5790-5799.
39. E. Weber, R. Gruetzner, S. Werner, C. Engler and S. Marillonnet, *PLoS One*, 2011, 6, e19722.
40. M. T. Miller, J. J. Higgin and T. M. Hall, *Nat. Struct. Mol. Biol.*, 2008, 15, 397-402.
41. H. T. Jenkins, R. Baker-Wilding and T. A. Edwards, *J. Struct. Biol.*, 2009, 167, 271-276.
42. Y. Wang, L. Opperman, M. Wickens and T. M. Hall, *Proc. Natl. Acad. Sci. U. S. A.*, 2009, 106, 20186-20191.
43. D. Zhu, C. R. Stumpf, J. M. Krahn, M. Wickens and T. M. Hall, *Proc. Natl. Acad. Sci. U. S. A.*, 2009, 106, 20192-20197.
44. G. Lu, S. J. Dolgner and T. M. Hall, *Curr. Opin. Struct. Biol.*, 2009, 19, 110-115.
45. Y. Y. Koh, L. Opperman, C. Stumpf, A. Mandan, S. Keles and M. Wickens, *RNA*, 2009, 15, 1090-1099.
46. L. Opperman, B. Hook, M. DeFino, D. S. Bernstein and M. Wickens, *Nat. Struct. Mol. Biol.*, 2005, 12, 945-951.
47. Y. K. Gupta, D. T. Nair, R. P. Wharton and A. K. Aggarwal, *Structure*, 2008, 16, 549-557.
48. C. Qiu, A. Kershner, Y. Wang, C. P. Holley, D. Wilinski, S. Keles, J. Kimble, M. Wickens and T. M. Hall, *J. Biol. Chem.*, 2012, 287, 6949-6957.
49. A. Galgano, M. Forrer, L. Jaskiewicz, A. Kanitz, M. Zavolan and A. P. Gerber, *PLoS One*, 2008, 3, e3164.
50. A. R. Morris, N. Mukherjee and J. D. Keene, *Mol. Cell. Biol.*, 2008, 28, 4093-4103.
51. J. P. Guilinger, V. Pattanayak, D. Reyon, S. Q. Tsai, J. D. Sander, J. K. Joung and D. R. Liu, *Nat Methods*, 2014, 11, 429-435.
52. T. Ozawa, Y. Natori, M. Sato and Y. Umezawa, *Nat Methods*, 2007, 4, 413-419.

53. J. Tilsner, O. Linnik, N. M. Christensen, K. Bell, I. M. Roberts, C. Lacomme and K. J. Oparka, *Plant J.*, 2009, 57, 758-770.
54. T. Yamada, H. Yoshimura, A. Inaguma and T. Ozawa, *Anal. Chem.*, 2011, 83, 5708-5714.
55. E. Querido and P. Chartrand, *Methods Cell Biol.*, 2008, 85, 273-292.
56. A. Cooke, A. Prigge, L. Opperman and M. Wickens, *Proc. Natl. Acad. Sci. U. S. A.*, 2011, 108, 15870-15875.
57. J. Cao, M. Arha, C. Sudrik, D. V. Schaffer and R. S. Kane, *Angew. Chem. Int. Ed. Engl.*, 2014, 53, 4900-4904.
58. J. Cao, M. Arha, C. Sudrik, A. Mukherjee, X. Wu and R. S. Kane, *Nucleic Acids Res.*, 2015, DOI: 10.1093/nar/gkv290.
59. M. Ghildiyal and P. D. Zamore, *Nat Rev Genet*, 2009, 10, 94-108.
60. R. Choudhury, Y. S. Tsai, D. Dominguez, Y. Wang and Z. Wang, *Nat Commun*, 2012, 3, 1147.
61. W. Zhang, Y. Wang, S. Dong, R. Choudhury, Y. Jin and Z. Wang, *Mol. Ther.*, 2014, 22, 312-320.
62. Y. Wang, C. G. Cheong, T. M. Hall and Z. Wang, *Nat Methods*, 2009, 6, 825-830.
63. T. Nakamura, Y. Yagi and K. Kobayashi, *Plant Cell Physiol.*, 2012, 53, 1171-1179.
64. A. Barkan and I. Small, *Annu. Rev. Plant Biol.*, 2014, 65, 415-442.
65. I. D. Small and N. Peeters, *Trends Biochem. Sci.*, 2000, 25, 46-47.
66. M. J. Howard, W. H. Lim, C. A. Fierke and M. Koutmos, *Proc. Natl. Acad. Sci. U. S. A.*, 2012, 109, 16149-16154.
67. R. Ringel, M. Sologub, Y. I. Morozov, D. Litonin, P. Cramer and D. Temiakov, *Nature*, 2011, 478, 269-273.
68. M. Takenaka, A. Zehrmann, A. Brennicke and K. Graichen, *PLoS One*, 2013, 8, e65343.
69. J. Ke, R. Z. Chen, T. Ban, X. E. Zhou, X. Gu, M. H. Tan, C. Chen, Y. Kang, J. S. Brunzelle, J. K. Zhu, K. Melcher and H. E. Xu, *Nat. Struct. Mol. Biol.*, 2013, 20, 1377-1382.
70. C. Shen, X. Wang, Y. Liu, Q. Li, Z. Yang, N. Yan, T. Zou and P. Yin, *Mol Plant*, 2015, 8, 667-670.
71. R. Barrangou, C. Fremaux, H. Deveau, M. Richards, P. Boyaval, S. Moineau, D. A. Romero and P. Horvath, *Science*, 2007, 315, 1709-1712.
72. J. E. Garneau, M. E. Dupuis, M. Villion, D. A. Romero, R. Barrangou, P. Boyaval, C. Fremaux, P. Horvath, A. H. Magadan and S. Moineau, *Nature*, 2010, 468, 67-71.
73. G. Gasiunas, R. Barrangou, P. Horvath and V. Siksnys, *Proc. Natl. Acad. Sci. U. S. A.*, 2012, 109, E2579-2586.
74. M. Jinek, K. Chylinski, I. Fonfara, M. Hauer, J. A. Doudna and E. Charpentier, *Science*, 2012, 337, 816-821.
75. F. J. Mojica, C. Diez-Villasenor, J. Garcia-Martinez and C. Almendros, *Microbiology*, 2009, 155, 733-740.
76. J. A. Doudna and E. Charpentier, *Science*, 2014, 346, 1258096.