DISSERTATION

METHODS FOR DETECTING AND DEVELOPING PROTEIN-PROTEIN OR PROTEIN-RNA INTERACTIONS

Submitted by

Brett D. Blakeley

Department of Chemistry

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Fall 2014

Doctoral Committee:

Advisor: Brian McNaughton

Alan Kennan
Nick Fisk
Melissa Reynolds
Olve Peersen

ABSTRACT

METHODS FOR DETECTING AND DEVELOPING

PROTEIN-PROTEIN OR PROTEIN-RNA INTERACTIONS

Potent and selective recognition of disease-relevant macromolecules – such as proteins and RNA – is the molecular basis of most pharmaceuticals . Historically, small (< 500 Da) molecules have filled this role. However, the overwhelming majority (~85%) of the proteome – and emerging therapeutic targets such as RNA – present a serious challenge to small molecule-dependent recognition. An alternative approach to potent and selective recognition and regulation of disease-relevant proteins and RNA is to use synthetic proteins. In contrast to small molecules, the size, relatively high folding energies (>10 kcal/mol) and functional group diversity (by virtue of proteinaceous amino acids) allow proteins to recognize – and potentially control – macromolecular receptors that evade small molecules. Presented here are two approaches to advancing the discovery of new proteins that recognize either disease-relevant protein or RNA targets. The first part of this thesis describes split superpositive GFP reassembly as a method to identify novel protein-protein interacting pairs in living cells (*E. coli*). The second part of this thesis describes basic studies to evaluate the suitability of a naturally occurring RNA Recognition Motif (RRM) as a scaffold for targeting disease-relevant RNA hairpins, and the development of new RRMs that target TAR RNA, a hairpin critical to HIV proliferation.

ACKNOWLEDGMENTS

I would first like to express my sincerest gratitude to my research advisor, Dr. Brian McNaughton. He has been an inspirational mentor, in science and many other areas of my life. Additionally, my graduate school experience was enriched by the peers I was lucky to share it with and to each of them, I send many thanks. I would never have had the confidence to pursue my goals if not for the unwavering encouragement from my family and for that I am forever grateful. I'd also like to thank my friends. Even long after I left the east coast, they continue to offer their unconditional support. In my accomplishments I also celebrate the life of my brother-in-law, Michael Knierim, a budding scientist in his own right. Finally, I would like to thank my most enthusiastic supporter, my best friend, my wife: SandiLynn. She has been my primary source of strength throughout my graduate studies, including the work presented here and, more importantly, the disappointment and despair in between. I acknowledge: I could not have done this without her.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

Aga	agglutinin protein
Ala or A	alanine
Asn or N	asparagine
Asp or D	aspartic acid
bp	base pairs
C9	cytosine at position 9 in the U1hpII hairpin loop
C10	cytosine at position 10 in the U1hpII hairpin loop
cfu	colony forming units
Cy5	cyanine dye number 5
Da	Daltons
ΔK50ΔM51	U1A with Lys50 and Met51 deleted
DHFR	dihydrofolate reductase
DNA	2′-deoxy-ribonucleic acid
dNTPs	deoxynucleotide triphosphates
dsDNA	double stranded DNA
EBY100	a strain of *S.cerevisiae* that constitutively express Aga1 for protein display
ELISA	enzyme linked immunosorbent assay
EMSA	electrophoretic mobility shift assay

FACS	fluorescence-activated cell sorting
FRET	fluorescence resonance energy transfer
Gal	galactose inducible promoter
G4	guanosine at position 4 in the U1hpII RNA hairpin loop
His or H	histidine
Ile or I	isoleucine
Leu or L	leucine
Lys or K	lysine
Met or M	methionine
mRNA	messenger RNA
MW	molecular weight
NMR	nuclear magnetic resonance
nt	nucleotide
PCA	protein complementation assay
PCR	polymerase chain reaction
Phe or F	phenylalanine
PPR	pentricopeptide repeat
PUF	pumilio / fem repeat proteins
RNA	ribonucleic acid
PPI	protein-protein interaction

PNA	peptide nucleic acid
RT	room temperature
RPM	rotations per minute
PUM1	human pumilio 1 protein
RRM	RNA recognition motif
SAR	structure activity relationship
Ser or S	serine
spGFP	superpositive green fluorescent protein
SPR	surface plasmon resonance
TAR	trans-acting response element
TBR_XG	TAR RNA binding RRM, from X^{TH} generation of sorting
tRNA	transfer RNA
Tyr or Y	tyrosine
U1A	human U1 snRNP protein
U8	uracil at position 8 in the U1hpII RNA hairpin loop
U1hpII	U1 hairpin II of the human snRNP complex
Val or V	valine

CHAPTER ONE

EXPANDING THE ROLE OF PROTEINS IN BASIC RESEARCH AND DRUG DISCOVERY: FROM PROTEIN-PROTEIN INTERACTIONS TO RNA RECOGNITION

1.1 Introduction

The pursuit of new drug leads and basic research tools includes a rich history of discovery that is centered on small organic molecules. These include notable achievements such as the development of β-lactams[1] as bactericidal agents, immuno-suppressants to facilitate organ transplants[2], morphine derivatives for pain management therapies, and microtubule binding agents for cancer therapy[3-5]. Generally, the desired biological activity of a small molecule drug or basic research tool hinges on a precise binding event within a buried hydrophobic pocket of a disease-relevant protein **Figure 1.1A**. Binding results in suppression or abrogation of disease-relevant protein function. Using this strategy, modern pharmaceuticals are able to precisely manipulate important classes of disease-relevant protein targets. However, severe limitations within this paradigm exist.

Small molecule-dependent regulation of protein function is inherently limited to protein targets with well-defined hydrophobic pockets amendable to small molecule binding **Figure 1.1B**. Thus, a small minority of the proteins in the human proteome (~20%) fit this criteria, and are largely limited to G-Protein Coupled Receptors (GPCRs) [6,7]. Thus, large swaths of protein

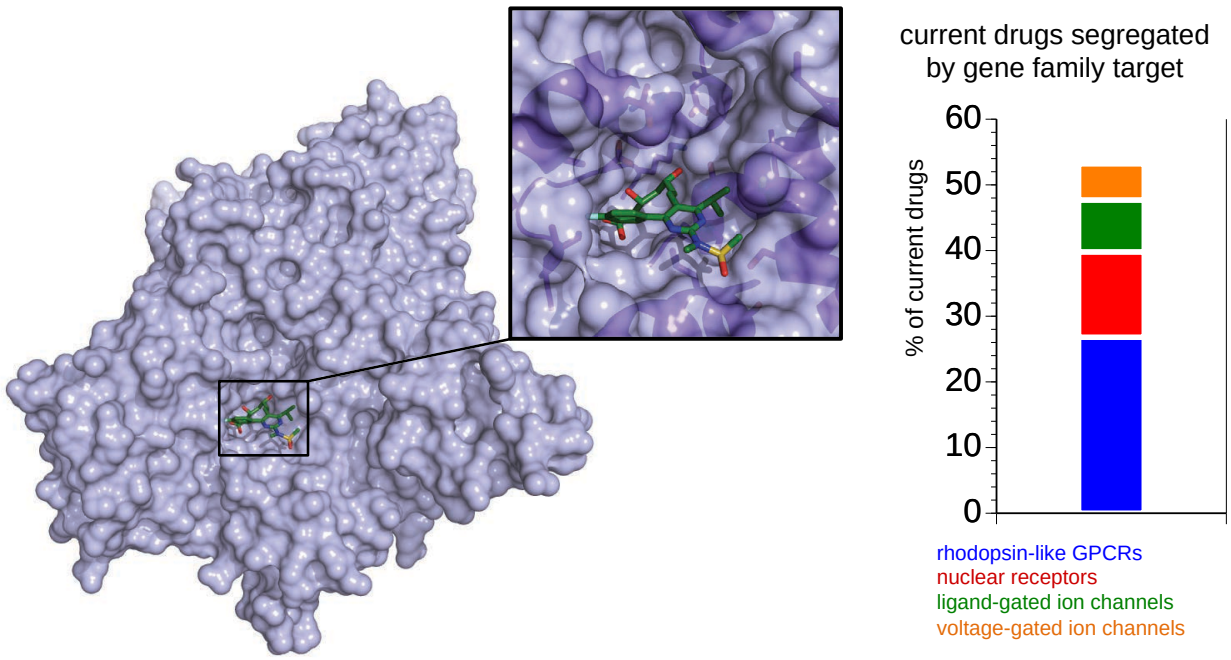


Figure 1.1 A Rosuvastatin (Crestor®) bound in a hydrophobic pocket of hHMG-CoA Reductase (PDB code: 1HWL) **B** stacked bar graph representing the current set of FDA approved drugs separated by their target protein class. Graph adapted from Overington, J.P.; Al-Lazikani, B.; Hopkins, A.L. *Nat. Rev. Drug Discovery* **2006**, *5*, 993.

space are resistant to, or entirely evade, small molecule control[8,9]. Moreover, small molecule-dependent modulation of emerging non-protein disease-relevant macromolecules, such as RNA, remains a daunting task.

Within the approximately 80% of the human proteome that is classified as "undruggable", proteins that engage in large protein-protein interactions are well represented[10-12]. Traditional organic compounds with a MW of < 500 Da lack the surface area and molecular complexity necessary to compete with many protein-protein interactions[13-15], which often involve expansive surfaces (typically ~1600 ±400 square anstroms[16]) **Figure 1.2A**. RNA also resists a traditional solution to molecular recognition[17]. The molecular framework of RNA is decorated with hydrophilic heteroatoms creating a complex interplay of electrostatic charges and hydrogen

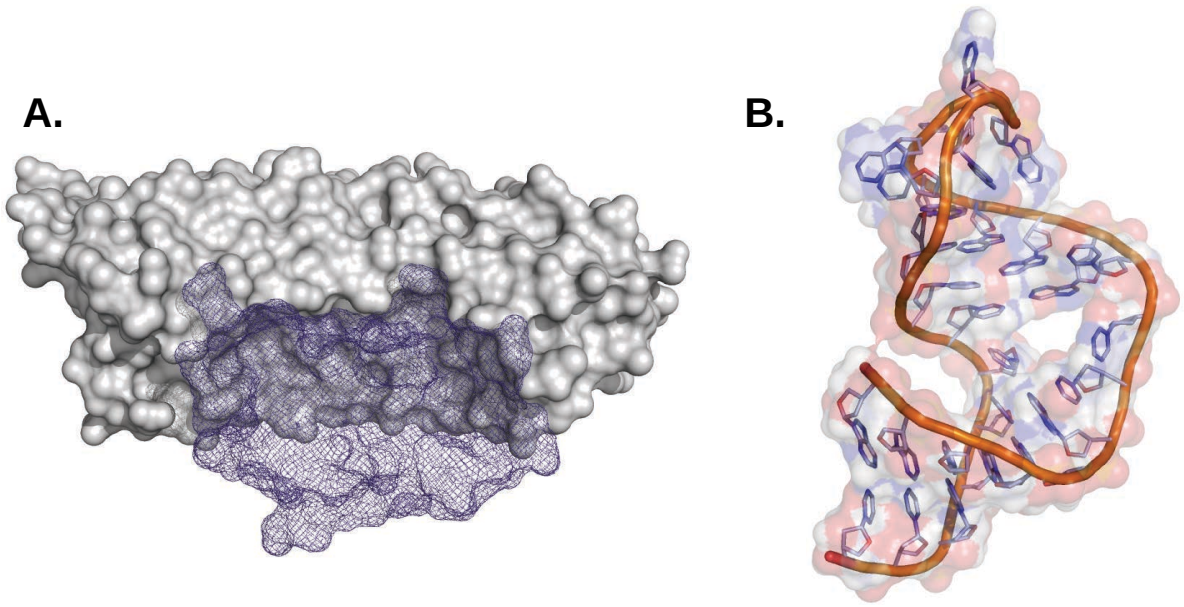


Figure 1.2 A The ankyrin repeat protein gankyrin (grey) binding to the S6 region of the 26S proteosome (purple) represents a large protein-protein interaction surface area that is difficult to disrupt using traditional small molecule drugs. (PDB code: 2DVW) **B** The structure of RNA presents a challenge to molecular recognition due to the low functional group diversity and complex electrostatic profile. (PDB code: 1UTS)

bonding interactions **Figure 1.2B**. Further complicating recognition, the tertiary structure of RNA tends to be more flexible and dynamic in aqueous solution relative to proteins or DNA, deterring the formation of static pockets or grooves commonly targeted by traditional drug discovery. The few small molecules that overcome these challenges to binding RNA tend to lack appreciable selectivity for a single RNA target, likely a result of the low functional group diversity among the nucleotide monomers that compose RNA[18,19]. Taken together, our inability to successfully target the overwhelming majority of disease-relevant proteins, and emerging targets such as RNA, require advances in the construction of new reagents that overcome the challenges of small molecule discovery, and in techniques that identify new interactions.

Part 1. Methods for the Identification of Protein-Protein Interactions

1.2 Quantitative Methods to Identify Protein-Protein Interactions

Many of the processes within the cell that maintain homeostasis are governed by protein-protein interactions, including inter- and intra-cellular signaling, protein degradation, metabolism, and apoptosis[20-22]. Critical to the basic study of human disease states are methods to identify protein-protein interactions within these intricate signaling networks. Additionally, recent trends in drug discovery have shown a surge of interest in biologics-based therapeutics[23-26], suggesting an increased emphasis on methods to identify new protein-protein interactions to develop proteinaceous drug leads.

A key component to understanding, or manipulating, protein-protein interactions is to identify regions or residues at the binding interface that are essential for the binding event to proceed. This is commonly achieved through structure-activity relationship (SAR) studies that measure how systematic mutations to the binding face affect the affinity of the interaction. This calls for reliable methods to quantify protein-protein interactions. Electrophoretic Mobility Shift Assay (EMSA) is an easy way to observe protein-protein binding interactions using equipment and reagents that are both low-cost and readily available in most chemical biology labs[27] **Figure 1.3A**. It also affords the added benefit of not requiring any chemical modifications to either interacting protein, though it can require a significant amount of the proteins of interest. This method works well for interactions with mid- to low-nanomolar dissociation constants but

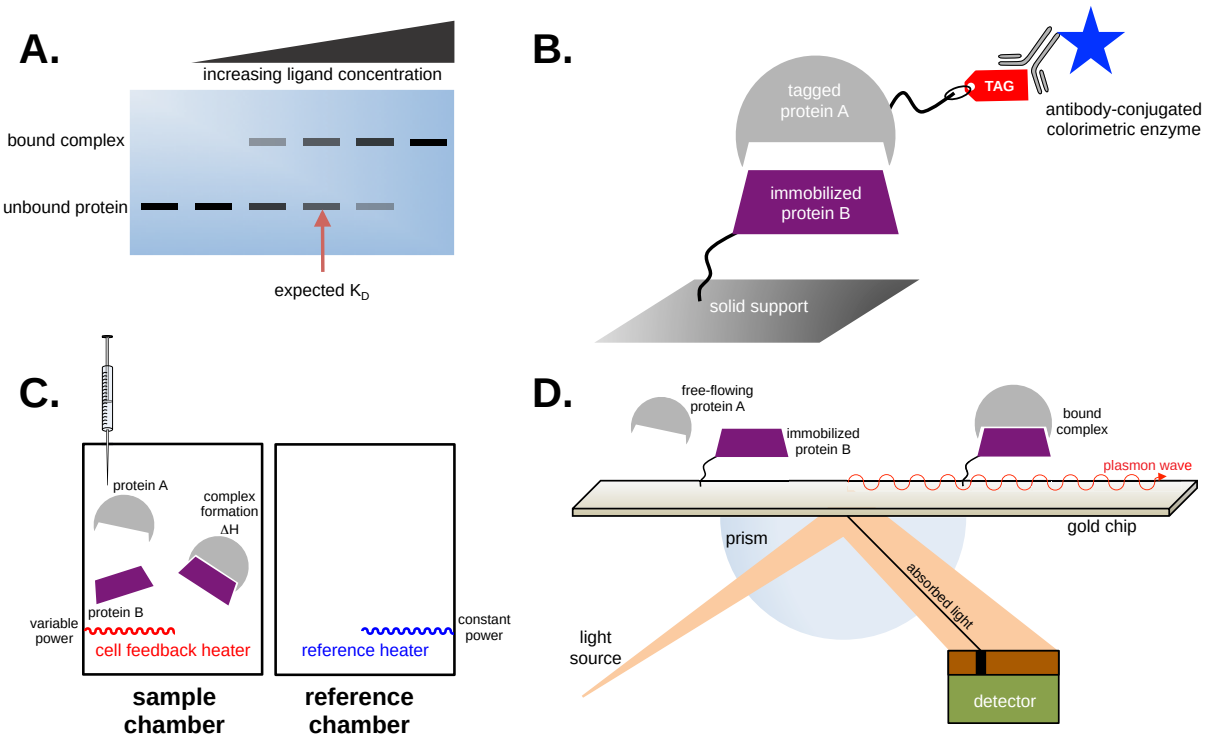


Figure 1.3 A General representation of an electrophoretic mobility shift assay. Increasing concentrations of protein A are added to a low and constant concentration of protein B. A shift in the protein band represents a bound complex. **B** Schematic of an enzyme linked immunosorbant assay. Protein B must be immobilized on a solid support. After incubation with tagged protein A and subsequent washing steps, an enzyme-conjugated antibody that binds the tag is introduced and the enzyme provides some colorimetric output. **C** Set-up of an isothermal titration calorimetry instrument. Protein A is injected through a syringe into a solution of protein B. Enthalpic changes from complexation are compensated relative to a reference chamber using a highly sensitive heating system. Output from the heater is recorded to derive thermodynamic binding data. **D** Schematic of a surface plasmon resonance experiment. A single wavelength light source is fed through a prism, giving a diversity of light angles. Changes in the state of the unbound or bound immobilized protein B on the gold surface correlate with the angle of light that excites plasmons on the gold surface.

decreases in efficacy for lower affinity interactions. With the proper instrumentation and planning, tens of different interactions can be evaluated in a reasonable time frame. An Enzyme-Linked Immunosorbent Assay (ELISA) requires more specialized and costly materials but is also a robust and reliable method for testing protein-protein binding interactions over a wide range of dissociation constants[28] **Figure 1.3B**. ELISA is typically used for qualitative or relative binding information of up to tens of binding interactions but can be quantitative[29] if paired with

a standard binding curve of a suitable control binding interaction.

Isothermal titration calorimetry (ITC)[30,31] **Figure 1.3C** and surface plasmon resonance (SPR)[32,33] **Figure 1.3D** are highly robust methods that require specialized dedicated instrumentation. As a result, these techniques demand a significant financial investment. Nevertheless, each of these techniques can be a very powerful tool for characterizing a protein-protein binding interaction. ITC is a solution-based method that does not require any modification to either binding partner and provides a full thermodynamic characterization of the system under examination. SPR does necessitate an immobilization step for one of the components, but is capable of determining both the association and dissociation rates of the interaction simultaneously. The impressive analytical power of these techniques is burdened by their slow workflow for assessing multiple binding interactions and, as mention, their high cost.

The increased availability and lower cost of instruments that monitor fluorescence has led to a rise in popularity of fluorescence-based methods for detecting interacting proteins and determining binding affinity[34-37]. These methods include fluorescence resonance energy transfer (FRET), fluorescence polarization, or fluorescence quenching. Most fluorescence-based methods will typically require chemical modification of one or both of the interacting proteins to incorporate the desired chromophore(s). A distinct advantage of fluorescent-based systems relative to those listed above is that they are amendable for high-throughput application for screening hundreds to thousands of potential protein-protein interactions in a tangible time-frame[38-40]. This can be of great utility when screening a large combinatorial population for potential binding partners.

1.3 High-Throughput Exploration of Protein-Protein Interactions

Identifying molecules with desired fitness within the immense chemical space represented by synthetic proteins requires robust methods to sort through functional and nonfunctional molecules[41-46]. Often this requires significant creativity and inventiveness on the part of the researcher with respect to experimental design. The exact parameters can vary depending on the target binding partner but most techniques fall into one of two general categories; selections, which link the binding event to a "life or death" decision mechanism in an organism, and screens, which endow some colorimetric output or otherwise observable phenotype that allows delineation of functional and nonfunctional proteins **Figure 1.4**. In a selection, identification of "winners" is operationally simple as the genotype of the surviving organisms can be verified using standard DNA cloning and sequencing techniques. In contrast, a screen requires the extra step of separating functional library members from those that are non-functional.

Display technologies such, as mRNA-, ribosome-, phage-, or yeast-display, are common screening techniques applied to identifying new protein-protein interactions[47-50 51,52]. The link between genotype and phenotype provided by display technologies enables facile identification of screened library members through DNA sequencing[53,54].

Tying the desired binding interaction to a fluorescent output makes deconvolution operationally simple through Fluorescence Activated Cell Sorting (FACS)[55-58] **Figure 1.5**. FACS instruments are commonly found in modern research facilities. They are capable of separating fluorescently labeled cells at a rate of ~10,000 events per second with high purity[59]. Although this technique is limited to sorting of micron-sized particles or cells, many common

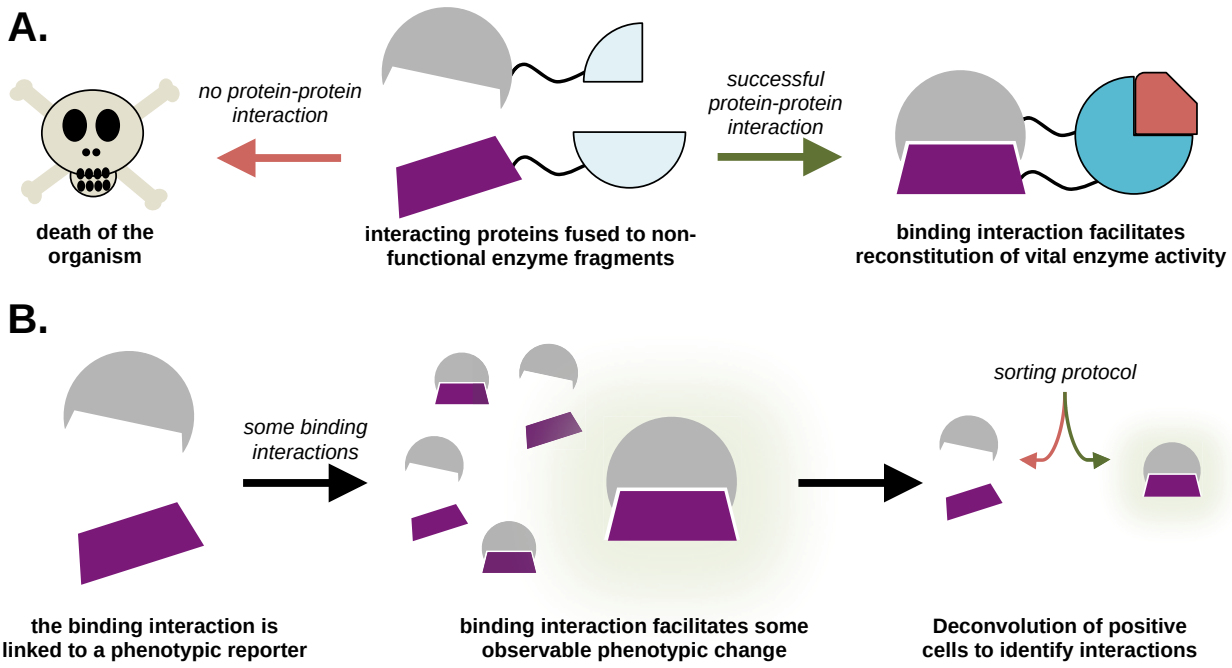


Figure 1.4 A Schematic of a selection for protein-protein interactions. The interacting proteins are linked to a "life or death" decision in the cell. No binding between these proteins leads to cell death (*left*). A binding interaction between the proteins leads to reconstitution of a survival pathway that rescues the cells, leading to their survival. **B** Schematic of a protein-protein interaction screen. The binding interaction endows the cells with a colorimetric phenotype so that interactions can be identified through some visual-based mechanism, usually applying specialized instrumentation.

screening technologies, such as fluorescent protein complementation, *in vitro* compartmentalization, and yeast display, have been designed to take advantage of this robust technique[60-62]. The incredibly high-throughput nature of FACS makes it a much more desirable option over other low-throughput methods for determining binding, such as yeast two- and three- hybrid assays, ELISA, or EMSA.

1.4 Observing Protein-Protein Interactions in Living Cells

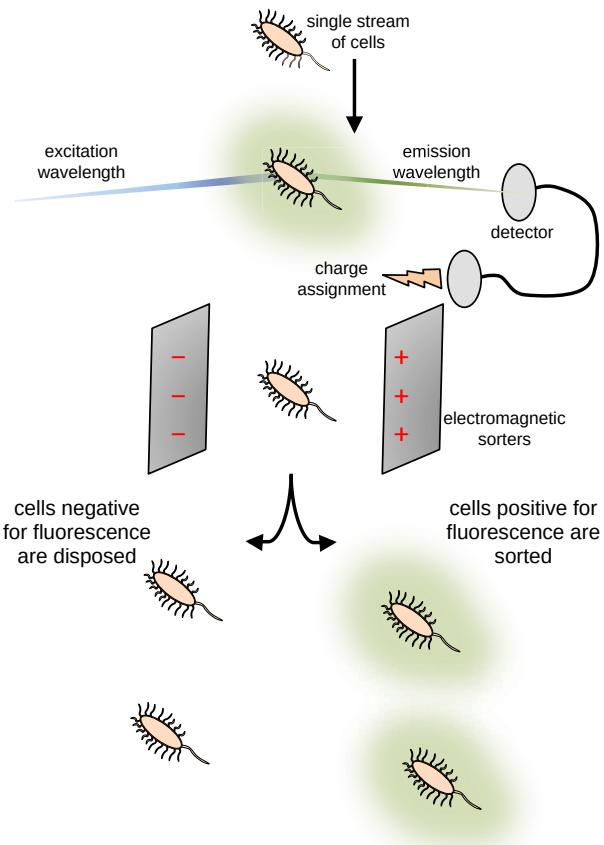


Figure 1.5 Schematic of Fluorescence Activated Cell Sorting (FACS). A network of tubing isolates single cells into a stream for analysis. As cells fall through the instrument, they are analyzed for fluorescence. Cells are then assigned a charge based on their fluorescence emission. Cells that are positive for the desired emission are sorted and kept while negative cells are discarded.

Many of the available methods to determine protein-protein binding are performed *in vitro* and therefore may not necessarily be relevant in a complex cellular environment. Some *in vivo* conditions can be replicated in an *in vitro* experiment, like temperature, but other variables, such as cellular localization or the presence of competing molecules, can be more difficult to account for. Ideally, the binding interaction is observed in a natural biological environment.

One of the first methods to address this need was the yeast two-hybrid assay[63-65]. As the name suggests, the interaction is tested in yeast, *S.cerevisiae*, which can function as a model system for higher eukaryotes. Unfortunately, the binary readout does not quantitate binding affinity and can be susceptible to a moderate rate of false positives. This method has largely fallen out of favor as more robust and higher-throughput systems for identifying protein-protein interactions have become available[66-68].

Another creative solution for monitoring protein-protein interactions in a cell is through

a protein complementation assay (PCA) [69-73]. Briefly, two non-functional fragments of a reporter protein are each fused to two interacting proteins of interest and expressed in a cell. Binding between these two proteins in the cytosol brings the non-functional fragments into close proximity, leading to reassembly of the fragments and reconstitution of the reporter protein activity **Figure 1.6**. A major benefit to assessing protein-protein interactions *in vivo* is that the observations are made in the presence of various biological metabolites, nucleic acids, and other proteins. This suggests a degree of selectivity for any observed interaction. Various colorimetric reporter proteins, such as β-lactamase, luciferase, and green fluorescent protein (GFP) **Figure 1.6** have been "split" for use in PCA[74-78]. Additionally, split-dihydrofolate reductase (DHFR) serves as a PCA-based selection by linking split-DHFR reassembly to a requisite step in the vital folate synthesis pathway[79,80]. The Michnick lab has used this assay to identify novel leucine zipper binding partners in *E.coli*[79], albeit with a moderate rate of false positives.

While PCAs tend not to give any quantitative affinity information, they can serve as a very powerful step in a workflow that also incorporates a more qualitative, low-throughput method. Both split-GFP and split-DHFR have been applied to high-throughput screening of combinatorial protein libraries to identify unique protein-protein interactions that can be characterized using other techniques[79,81]. The speed with which new protein-protein interactions can be identified and the inherent biological relevance of those interactions make PCAs an attractive option for researchers.

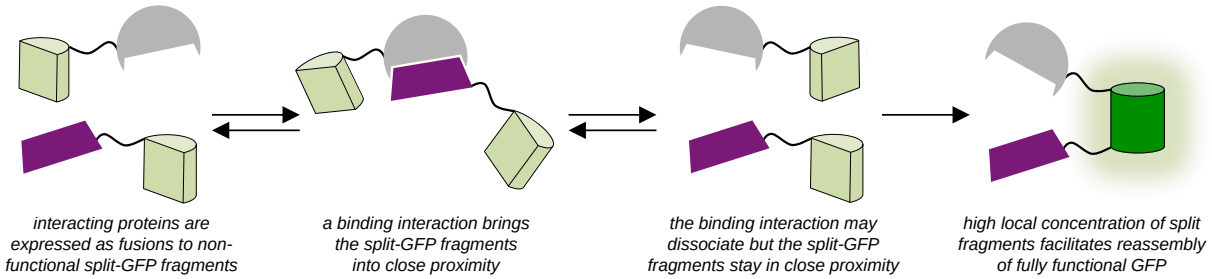


Figure 1.6 Schematic of the protein complementation assay, split-GFP. A binding interaction between the proteins of interest (grey and purple) leads to a high effective molarity of the non-functional fragments of GFP, driving fragment reassembly (far right).

1.5 The Virtues and Limitations of Split-GFP-Based Screens for Protein-Protein Interactions

The split-GFP system was pioneered by Lynn Regan's lab at Yale, initially to monitor the interaction of leucine zipper peptides *in vivo*[76,82]. They have applied this method to combinatorial screening of a protein library to find tetraticopeptide repeat proteins with new binding specificities for a particular α-helix peptide[81]. Successful protein-protein binding interactions are indicated by fluorescence from reassembled GFP molecules in *E.coli* cells. Positive cells are then collected using FACS to enabled the rapid analysis of a ~10^8 member library in under three hours.

A valuable benefit of split-GFP with respect to high-throughput screening is a low to zero false positive rate, however, there is thought to be a non-zero rate of false negatives deriving from the aggregation of the split protein fragments in the aqueous cellular environment[83] **Figure 1.7A**. Perhaps it is to be expected, given the β-barrel tertiary structure of GFP, that, upon splitting the protein and exposing the hydrophobic core of the barrel, the two fragments will have limited solubility in aqueous solution. As a result, a poor signal to background ratio is

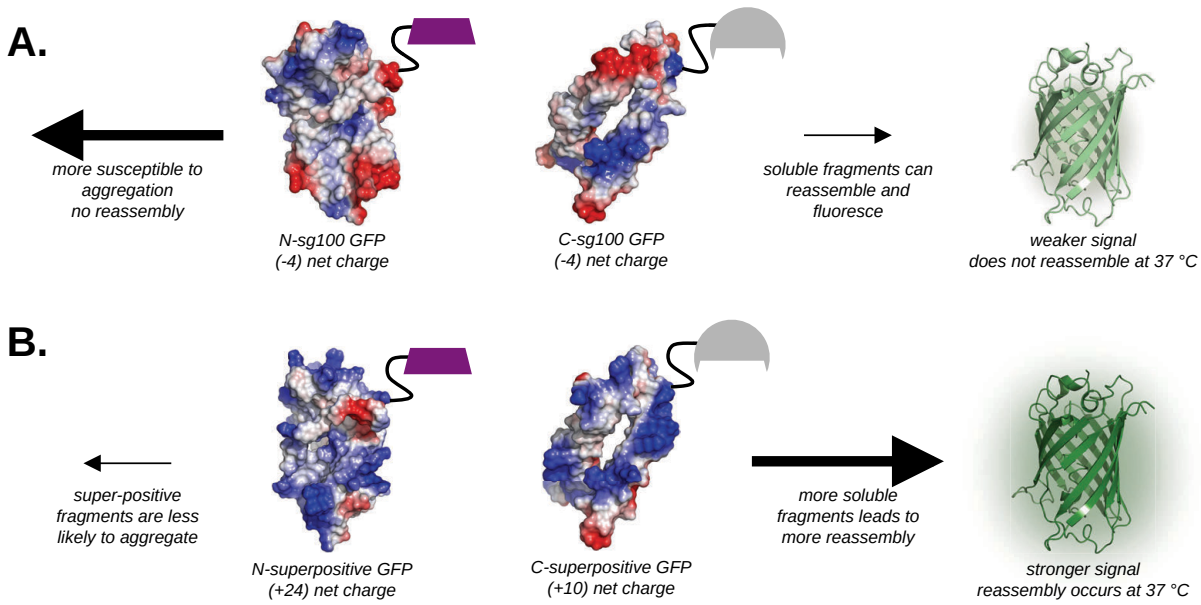


Figure 1.7 A The original split-GFP fragments, from sg100GFP, have a low theoretical net charge and a large amount of exposed hydrophobic surface area. Aggregation of the fragments leads to poor performance of the assay, decreasing its usefulness. **B** Superpositive variants of the split-GFP fragments have a higher theoretical net charge, allowing them to resist aggregation. Within, we demonstrate the resulting increase in performance of the split-superpositive GFP assay, including the ability to monitor protein-protein interactions at physiological temperature. (PDB code: 2B3P)

observed and assays cannot be conducted above 25 °C. This dilutes the advantage of *in vivo* analysis as protein-protein interactions likely behave much differently at 25 °C than at physiological temperature (37 °C). This limitation also frustrates screening efforts by curbing the rate at which hits can be identified from a combinatorial population.

Reported here is an effort to redeem the virtues of the split-GFP assay by improving the solubility of the split fragments, which would rescue functional members of a library that may otherwise be lost to aggregation **Figure 1.7B**. Supercharging a protein through surface mutations that increase its theoretical net charge has been shown to endow proteins with desirable properties including increased stability and solubility[84]. We reasoned that applying this concept to the fragments of split-GFP might correct their poor solubility and improve the

performance of the assay.

What we found was that split-superpositive GFP gave more than an order of magnitude brighter fluorescent signal than the traditional split-GFP assay in side by side comparisons[85]. The superpositive variant also produced an observable signal over background in a much shorter time, requiring only three hours of induction before analysis, as oppose to an overnight incubation. Perhaps most impressively, the improved assay was able to report binding of interacting proteins at 37 °C, thus enabling analysis of protein-protein interactions at physiological temperature. This is a welcome improvement over the 25 °C limit of the original assay. The higher temperature establishes more stringent conditions for library screening and greatly increases the chances that newly identified protein-protein interactions are relevant in a natural cellular environment. We also demonstrated that the positioning of the interacting proteins relative to the split-superpositive fragments is irrelevant for GFP reassembly, allowing us to propose a unique model for GFP reassembly.

Part 2. Potent and Selective RNA Recognition as a Means for Drug Discovery

1.6 Nucleic Acids Present Compelling Therapeutic Targets

While traditional methods of modulating biological function have focused on targeting proteins, an alternative approach might imposed control at the transcriptional or translational level by binding the nucleic acids that code for a target protein **Figure 1.8**. One could envision a mechanism by which a reagent specifically binds to and blocks the promoter region on double stranded genomic DNA upstream of a gene of interest. Other potential regulation points include RNA processing, mRNA translation, or non-coding RNA function. Thoughtful execution of these concepts would require molecular recognition reagents that potently and specifically bind to targeted nucleic acids. Examples from the literature suggest that developing molecular recognition reagents for nucleic acids is a daunting task, but holds great potential for unique methods of biochemical manipulation and treatment of disease[86-95].

The polyamide series of compounds developed and characterized in Peter Dervan's lab serve as a seminal achievement in molecular recognition of nucleic acids[96-99]. Dervan and coworkers performed exhaustive structure activity relationship studies on the binding interaction between distamycin and it's double stranded DNA binding partner[100] to determine the rules that govern the affinity and specificity of this interaction. Thus using distamycin as a

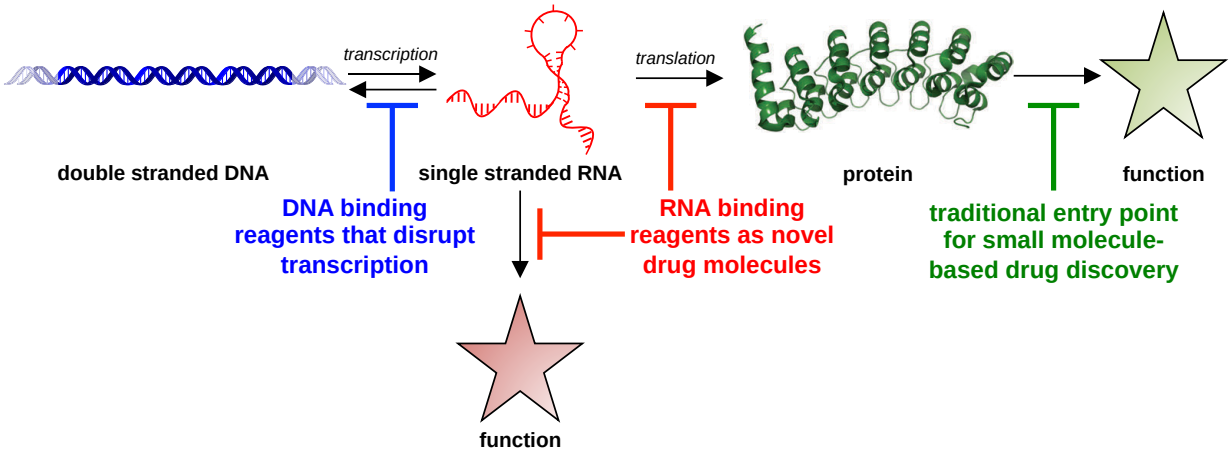


Figure 1.8 The central dogma of biology with potential entry points for synthetic reagent-mediated control over biological processes highlighted. The traditional model for drug discovery heavily favors the targeting of protein function to enact control over biological processes (green). This paradigm ignores the possibility of developing synthetic reagents that target nucleic acids in an effort to control DNA transcription (blue), RNA translation (red), or other important RNA functions (red). (PDB code: 2DVW)

template scaffold, they were able to identify monomer units that make specific electrostatic contacts with nucleobase functional groups in the minor groove of a DNA double helix. Each of their amide monomer units has a preference for binding either a G:C, C:G, A:T, or T:A base pairing **Figure 1.9**, which provides a degree of selectivity for recognizing a specific tract of double stranded DNA up to four base pairs. This selectivity can be programmed into the synthesis of a given polyamide compound for targeting any strand of DNA[101]. Recent studies investigating the therapeutic utility of these molecules suggest they are able to outcompete certain transcription factors for DNA binding sites[102-104]. Collectively, this work is an example of the requisite paradigm shift for advancing beyond the current model for drug discovery and reaching for "undruggable" targets.

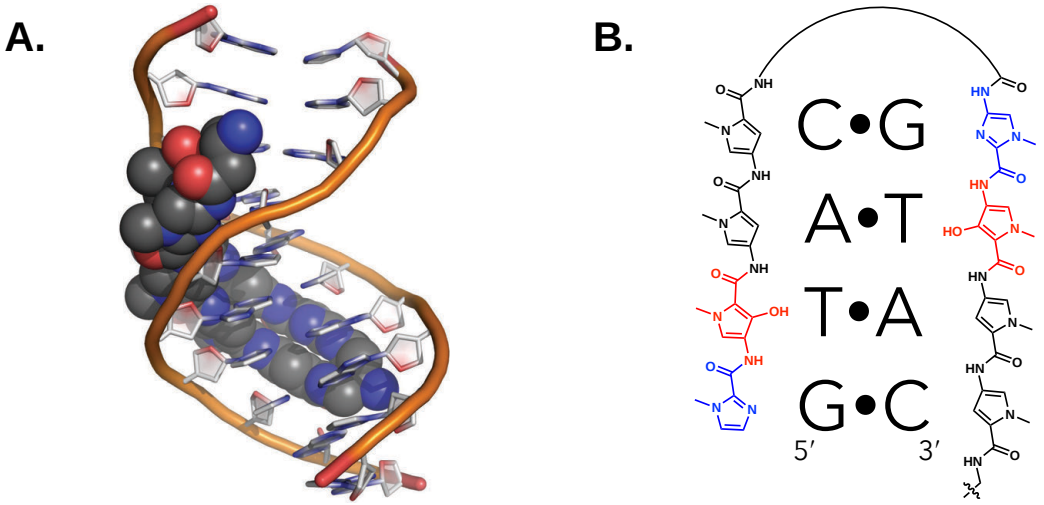


Figure 1.9 A Crystal structure showing a polyamide compound binding in the minor groove of a tract of double stranded DNA (PDB code: 3OMJ) **B** The code for recognition of double stranded DNA by the amide monomers of Dervan's polyamide compounds. Figure was adapted from http://dervan.caltech.edu/polyamides.php

1.7 The complex structure of RNA *in vivo* frustrates molecular recognition efforts

Molecular recognition of DNA certainly presents a challenging problem that has, essentially, been elegantly solved by Dervan's polyamide compounds, but in comparison, the molecular recognition of RNA seems to be a much more sophisticated problem that will likely require a more sophisticated solution. Generally, RNA exists *in vivo* as a single stranded polymer that adopts a wide range of secondary structures and can sample numerous conformations with thermally available energy under physiological conditions[105-108]. This is in contrast to the rigid and static double helical structure of DNA *in vivo*, where this single shape of DNA requires only a single chemical scaffold for recognition **Figure 1.10**. An analogous platform for RNA recognition likely calls for multiple scaffolds, each tailored to a specific RNA secondary structure

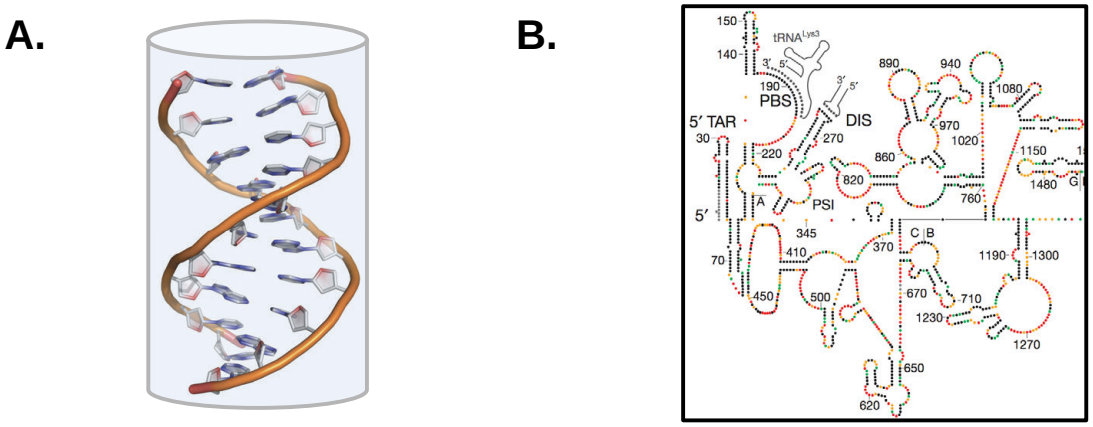


Figure 1.10 A The structure of DNA *in vivo* is relatively static and constant, forming the anti-parallel double helix, that, in a simplistic sense, represents a long cylindrical shape (PDB code: 3OMJ) **B** The structure of RNA *in vivo* is typically much more dynamic as it primarily exists in the single stranded form and adopts a wide range of secondary and tertiary structures, as seen here in a 5′-section of the HIV-1 genome measured using SHAPE. Adapted from Watts, J. M. *et al. Nature*, **2009**, *460*, 711-716.

or shape.

In search of new ligands that bind and, by extension, disrupt the function of specific RNA targets, early efforts centered on small organic molecules that obey the sacrosanct Lipinski's rule of five[17,19,89,109-111]. However, shoehorning RNA into the protein-targeting infrastructure of traditional drug discovery has since born out little success. A major component in achieving negative ΔG in a protein-small molecule interaction is driven by entropy from displaced water molecules at the binding site and enthalpic contributions from burying hydrophobic surfaces in the aqueous environment[112]. The increase in disorder from newly freed water molecules greatly outweighs any entropic cost of restricting degrees of freedom in the ligand and at the binding site on the protein. This principle does not generally translate to RNA binding due to its highly dynamic nature in aqueous solution. In contrast to the relatively limited number of lowest energy conformations adopted by globular proteins, single stranded RNA adopts a wide variety

of structurally diverse conformations[113-115]. Therefore locking an RNA binding pocket into place with a rigid small molecule demands a higher entropic penalty. That penalty can be paid by a larger ligand that displaces more water molecules; For example: high affinity protein-RNA interactions that encompass large surface areas[116,117]. Additionally, the monomeric units of RNA are relatively uniform in chemical composition and shape relative to the amino acid residues that make up proteins. Achieving specificity for an RNA ligand would likely require numerous contacts over large distances to delineate between discrete RNA targets. Collectively, these factors conspire to confound "small" molecule reagents developed for RNA recognition

1.8 Non-Ribosomally Derived Molecular Scaffolds Applied to Binding RNA Targets

A "privileged scaffold" generally refers to a molecular framework that has a proclivity to recognize a certain shape or class of biopolymer [118,119]. The notion has been successfully applied to protein-targeted reagents by populating combinatorial small molecule libraries with "privileged scaffolds" and test them for biological activity[120-123]. One of the first privileged scaffolds identified for binding RNA is the molecular framework of 2-deoxystreptamine **Figure 1.11** *left*, which serves as the backbone for the aminoglycoside class of natural products[124-126].

Naturally occurring aminoglycosides, including anamycin and neomycin **Figure 1.11**,

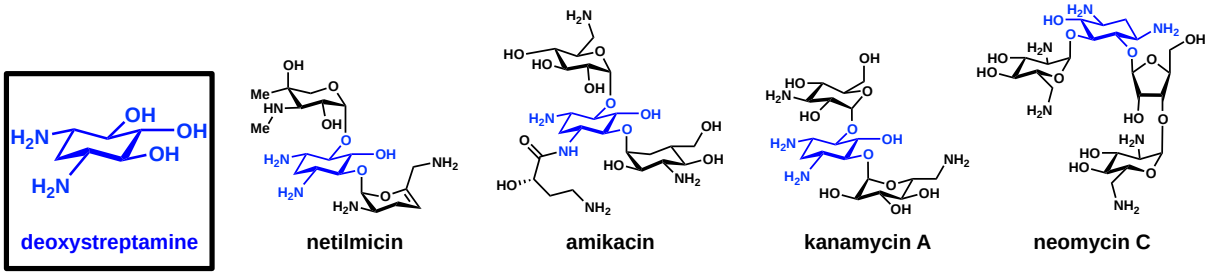


Figure 1.11 The deoxystreptamine scaffold serves as the core for the aminoglycoside class of RNA binding compounds. Many of these compounds were initially isolated from various bacterial species (kanamycin and neomycin). Synthetic chemists have since explored the chemical space centered on this molecular scaffold to find non-natural RNA binding molecules with anti-bacterial activity (netilmicin and amikacin).

are used to treat certain types of prokaryote infections in mammals[127, 128]. Exploration of chemical space within this privileged scaffold has led to non-natural synthetic variants of aminoglycosides[129,130]. For example, netilmicin and amikacin **Figure 1.11** are synthetically prepared and also act as antibiotic reagents for staving off infections from some multiple-drug resistant strains of pathogenic bacteria[131-133]. Thus, chemical diversification of a privileged scaffold can lead to new binding properties and biological activities.

Limiting widespread application of aminoglycosides as universal RNA targeting ligands is the glaring and inherent caveat of "small" molecules is that they are small, as described above[134-136]. While the ribosome is a relatively static globular assemblage of proteins and RNA, most other naturally occurring RNAs of interest *in vivo* are innately more dynamic, demonstrating higher degrees of freedom.

Recently, researchers have branched out to explore more complex chemical structures beyond the ~500 Da limit traditionally imposed[109] upon small molecules, perhaps inspired bythe success of Dervan's DNA binding compounds **Figure 1.12**. The Hergenrother group pioneered a linear progression toward larger RNA ligands by linking together two

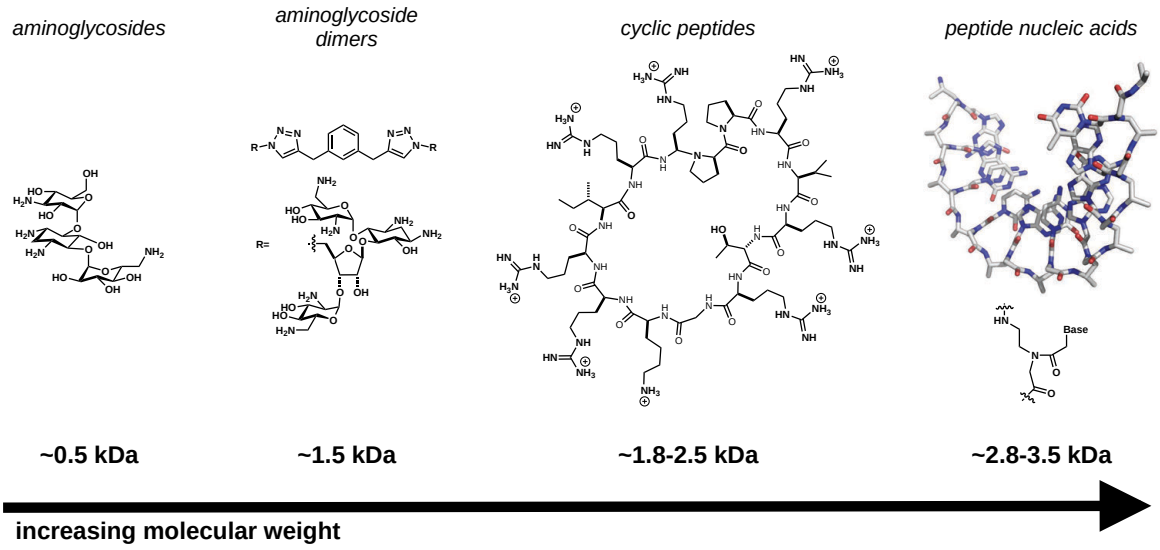
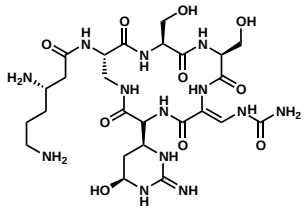


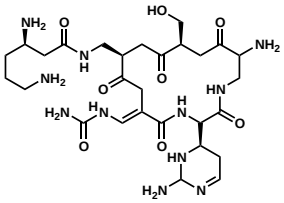
Figure 1.12 Representative molecules of common scaffolds in RNA binding compounds. Depicted here is a trend of increasing molecular weight in recently explored molecular scaffolds for targeting RNA. (PDB code: 3PA0)

aminoglycosides with a rigid linker to make aminoglycoside dimers[137,138] **Figure 1.12** *center left*. These compounds are slightly larger than 1,000 Da and bind RNA with low nanomolar affinity, but still lack appreciable target selectivity[139]. Disney and coworkers have more recently attempted to address the issue of poor target selectivity by developing a 2 × 2 library screening method of pseudo polymeric molecules composed of aminoglycosides and Hoechst stain[140-142]. Using this technology, they have identified compounds that disrupt a protein-RNA interaction causative of Multiple Sclerosis *in vitro*[143-145].

Cyclic peptides are another class of Nature-inspired macromolecules (~0.5-1.5 kDa) that have garnered attention in the RNA-targeting field[146-149]. Similar to the aminoglycosides, viomycin and capreomycin **Figure 1.13** demonstrate bactericidal activity, but are structurally unique from aminoglycosides[150,151]. These compounds have found use as treatments for muliple-drug resistant strains of tuberculosis[152,153]. Peptides that are conformationally constrained



viomycin



capreomycin

Figure 1.13 The structures of two naturally occurring cyclic peptides with bactericidal activity. Although their structure is unique from aminoglycosides, their mode of action is similar in that they bind to the bacterial ribosome and disrupt protein translation.

through covalent "staples" or cyclization have been shown to enhance binding properties and biostability relative to linear peptide equivalents[154-157]. Synthetic reagents that disrupt protein-protein interactions have been generated by isolating the binding moiety of a protein and synthesizing that region as a constrained peptide[157,158]. A number of researchers have also reported efforts toward applying the desirable traits of conformationally constrained peptides to developing new RNA ligands[154,159-161].

Varani and coworkers sough to develop a constrained peptide mimic to study the conformational changes of the HIV-1 Trans-Acting Response element (TAR) RNA hairpin in bound and unbound states[148,159,162-164]. They synthesized cyclic peptide mimics of the TAR RNA binding region from the HIV-1 *tat* protein. Using NMR, they identified contacts made between a cyclic peptide and both the hairpin loop and the stem loop bulge of the TAR hairpin, which are separated by a distance larger than the reach of most small molecules[160] **Figure 1.14A and B**. They also noted a conformational change in the TAR hairpin that prevented the full length *tat* protein from binding but did not activate translation of downstream viral proteins[163], suggesting the cyclic peptide mimetic may be of therapeutic relevance. This and other RNA-binding cyclic peptides[165-167] mimic a known binding interface, which places significant limitations on this approach.

Neilsen and coworkers appropriated an alternative bio-mimicking strategy by reconciling

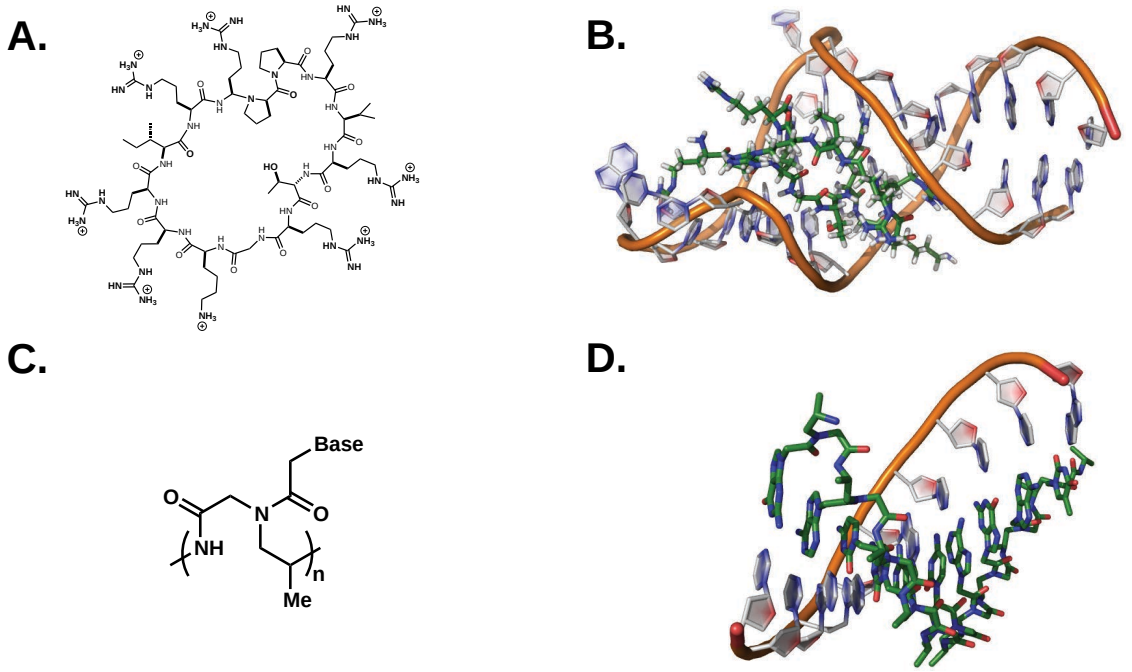


Figure 1.14 A Structure of a cyclic peptide that binds to the TAR RNA hairpin **B** Crystal structure of the cyclic peptide (shown with green carbons) bound to the TAR RNA hairpin (PDB code 1MNB). **C** Shown is the monomer unit of a common peptide nucleic acid backbone **D** A crystal structure of a peptide nucleic acid bound to it's complementary strand of DNA. (PDB code: 3PA0) Since these molecules bind through traditional Watson-Crick type interactions, they have also been used to target RNA.

the build-in Watson Crick base pairing of nucleic acids with the *in vivo* stability of amide bonds[168-172] **Figure 1.14C and D**. These peptide nucleic acids (PNAs) can weigh up to 3.5 kDa and are able to achieve good binding dissociation constants while maintaining excellent bioavailabilities and half lives. In aggregate, the representative examples above show that RNA ligands with increasing molecular weight have provided new synthetic compounds of bio-mimetic origin that expand the scope of biological probes available to the chemical biology community and have generated new potential drug leads.

1.9 Applying Synthetic Proteins to Drugging the "Undruggable"

The forces of evolution have molded a tremendous diversity of proteins over billions of years with varying functions throughout the human body. Researchers have begun exploiting the principles of natural evolution, by synthetically manufacturing diversity and manually influencing selective pressures in a laboratory setting[42,173-177] **Figure 1.15**. Expanding the functional repertoire of proteins beyond the natural set has vast implication for reinventing modern drug discovery. For example, natural proteins have evolved to be "just good enough" for their assigned function; however, Nature has not selected high affinity ligands for many of the potential biopolymer targets that may be of interest for their therapeutic or general research implications. Additionally, proteins boast a highly diverse set of chemical functional groups that aid in facilitating productive contacts with potential ligands. These functional groups are generally positioned in a dense and spatially defined arrangement on the amide backbone scaffold.

Lastly, the chemical composition and molecular framework of a protein can be quickly and precisely modified using broadly available laboratory methods (saturation mutagenesis, error prone PCR, DNA shuffling, other various PCR based methods, etc.) so that many different variations of an original protein template can be generated with relative ease[178,179]. The scale and speed with which chemical diversity can be integrated into a selected protein scaffold is simply unmatched when compared to analogous molecular scaffolds typically applied to biopolymer recognition such as small organic compounds or non-ribosomally derived biomimetics.

Intriguing findings have been reported by groups examining synthetic proteins (proteins

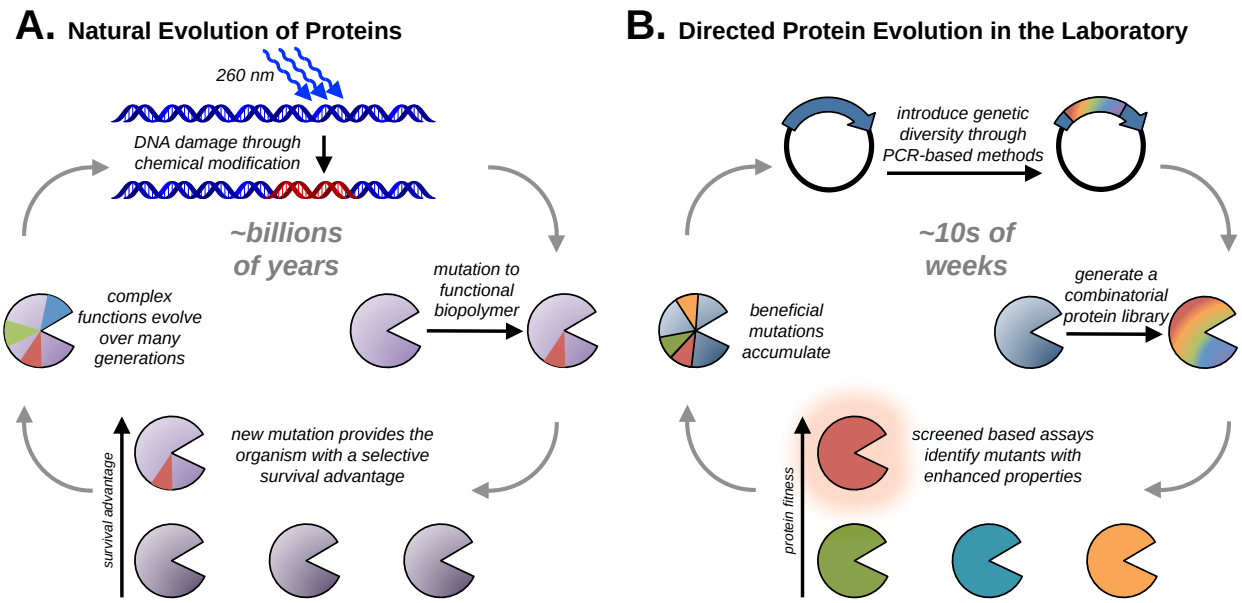


Figure 1.15 A Alterations to the genetic information in an organism can occur through exposure to UV irradiation, free radical damage, or some other chemical modification. If this damage occurs in the open reading frame of a protein, mutation may, in rare cases, cause an improvement or a change in function of the gene product. Beneficial mutations provide a survival advantage to the organism, allowing the improved genetic information to be selected and propagated through future generations. This process takes place over extremely long periods of time, not practical for research applications **B** Numerous PCR-based methods allow researchers to expediently diversify a gene of interest in a spatially controlled manner. The resulting combinatorial population of protein products is then assayed for the desired function, screening out highly functional members for incorporation into subsequent rounds. This process can be carried out in a much more tangible time frame in a research setting.

not found in Nature) as a possible solution to the RNA recognition problem. One particularly notable example is the engineered Pumilio/*fem-3* (PUF) repeat proteins[180-182]. PUF proteins consist of repeating α-*helix-turn-* α *-helix-turn-* α *-helix-turn* motifs that align in a crescent shape. Helices on the concave face of the protein present residues that participate in π-stacking, π-cation or Watson Crick-type interactions with ~seven bases of a single stranded RNA binding partner **Figure 1.16**. This elegantly simple binding mode endows PUF proteins with very high affinity (K_D as low as ~50 pM) and excellent selectivity.

The Tanaka Hall group, in collaboration with the Zamore group, were able to decipher

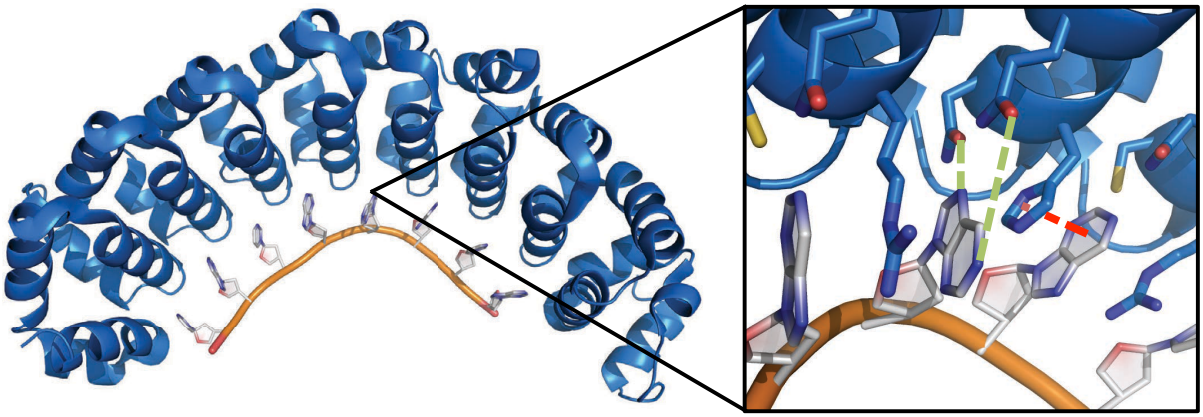


Figure 1.16 Representative member of the pumilio repeat protein class of RNA binding proteins. This class recognizes single stranded RNA through well positioned Watson Crick-type hydrogen bond interactions (*right* shown in green) between amino acids and nucleotides, as well as π-stacking interactions (*right* shown in red) between positively charged residues and the aromatic rings of the nucleotides. (PDB code: 3Q0Q)

the PUF code for recognition of RNA, allowing them to reprogram the sequence selectivity of human Pumilio1 (PUM1)[183,184] to bind a single stranded RNA with little sequence homology to the cognate binding partner of wt PUM1 while demonstrating tight binding affinity, on par with the native interaction. Further studies have shown that this strategy is broadly applicable for any single stranded RNA. Tanaka Hall and coworkers have since leveraged redesigned PUF scaffolds to manipulate RNA pathways *in vivo*[185]. Similar work on Pentatricopeptide Repeat Proteins (PPR) by Small and coworkers also demonstrated a programmable RNA recognition scaffold[186-188]. Together, these scaffolds serve as a general platform for single stranded RNA recognition reagents that can be generated based on a predetermined set of rules, analogous to Dervan's polyamide compounds.

1.10 Fitness Landscapes and Scaffold Selection

As researchers explore new and uncharted protein space, it is apparent that the great majority of this space contains unfolded and/or nonfunctional structures. Our current understanding of how primary protein sequence informs tertiary structure, and ultimately function, is foggy and unreliable. Francis Arnold's representation of a protein's plotted "fitness landscape"[189] suggests that functional structures are clustered together in sequence space and that these "hills" of fitness can be ascended by incremental changes to the structure, or mutations to the sequence[190,191]. We can think of these clusters as each representing a "privileged" protein scaffold for the desired function. It would obviously be prudent to begin an exploration through protein sequence space by focusing on this privileged scaffold area, thus starting half way up a fitness hill.

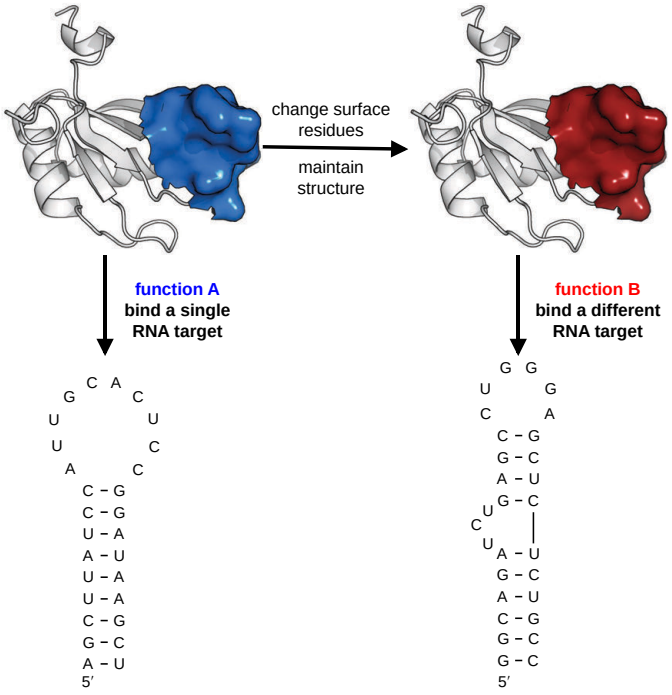


Figure 1.17 General representative scheme of protein semi-design. Regions of the protein known to be critical for recognition of an RNA binding partner are modified to facilitate recognition of a new binding partner. (PDB code: 1URN)

A privileged scaffold can derive from a known protein with a function tangential to the desired task. In this way, the scaffold can be essentially "redirected" to perform a similar, but new and desirable function instead of building a completely new structure from *a priori* principles **Figure 1.17**. For example:

reprogramming an RNA binding protein to bind an RNA target that is of therapeutic interest preferentially over it's native RNA binding partner. To do this, one must also carefully consider how amendable a scaffold might be to mutations in terms of stability. Ideally, a candidate protein will accommodate mutations, usually at the active site, that alter the inherent function of the protein without disrupting stability or overall structure[192]. Protein scaffolds that are easily moldable in terms of activity, or highly plastic scaffolds, are best suited for protein reengineering efforts.

1.11 Developing a Novel Platform for Targeting Folded RNAs

Taking a bird's eye view of the RNA-targeting landscape, we can begin to see multiple approaches developing and maturing, each utilizing unique molecular platforms that allow researchers to generate new recognition elements for a desired class of RNA **Figure 1.18**. This is to further suggest that the broad structural diversity of functional RNAs *in vivo* does not lend itself to a "catch all" molecular recognition solution, as seen with Dervan's DNA binding polyamides. At the genesis of this burgeoning field, many of the strategies described here are built on a basal level understanding of the requirements for generating new ligands for RNA. For example: anti-sense technologies, such as PNAs[193], rely on the Watson-Crick base pairing rules, elucidated over half a century ago[194]. The next generation of synthetic RNA binding compounds should, instead, appreciate and account for the three-dimensional structure inherent to the function of a discrete folded RNA target.

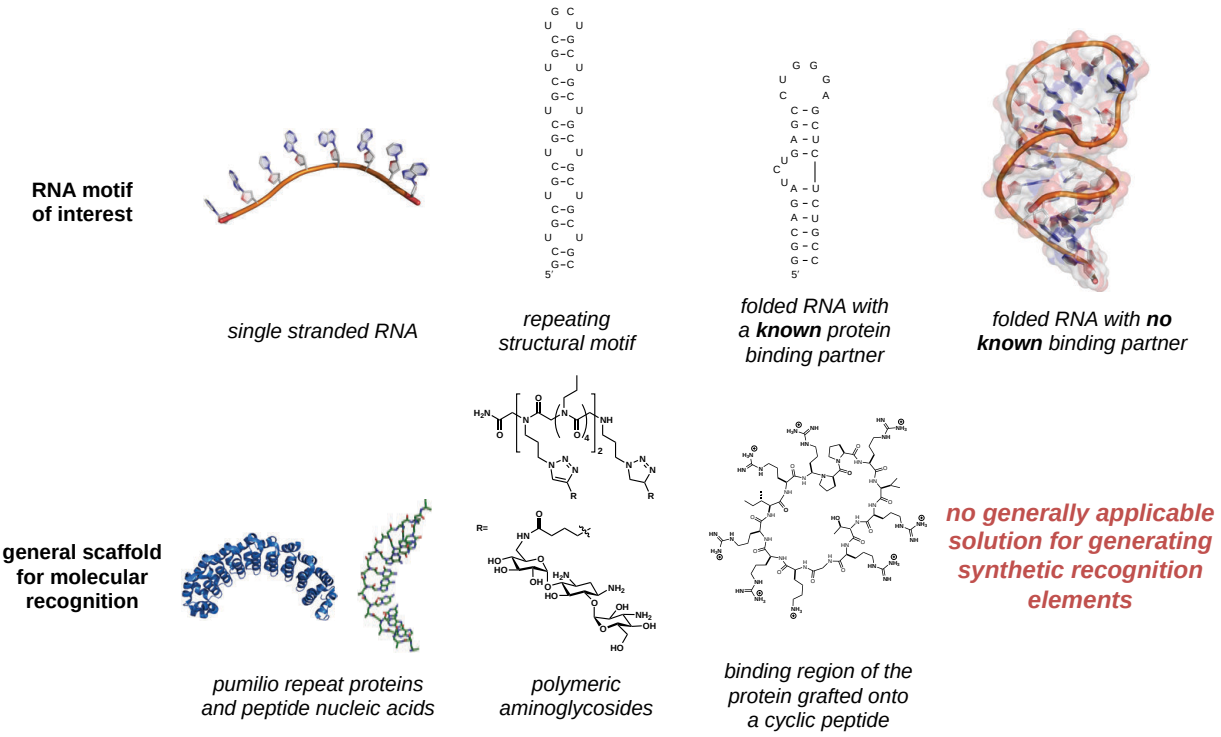


Figure 1.18 The myriad of three dimensional structures that are represented in physiologically relevant RNA molecules calls for unique molecular scaffold solutions for recognizing different RNA targets. Despite the tremendous progress highlighted in this table, many folded RNA structures continue to elude a generally applicable approach to a molecular recognition. (PDB codes: left to right 3Q0Q, 3PA0, and 2LDL)

The function of physiologically relevant RNA moieties is intimately tied to their tertiary structures[106,107], suggesting that recognition of folded RNAs is necessary for applications toward broadly manipulating RNA biochemistry. RNA binding proteins in Nature routinely identify RNAs by their defined tertiary structure[117]. In this work, we borrow one of these protein scaffolds, the RNA Recognition Motif (RRM)[195-197], as a template for a semi-design approach to new RNA binding reagents **Figure 1.19**. This strategy serves as an alternative to chemical diversification of aminoglycosides or tuning the RNA binding selectivity of the pumillio repeat protein scaffold. In a similar manner we look to modify the chemical identity of an RNA-binding scaffold, but one that recognizes folded RNAs **Figure 1.19**. This approach should allow new

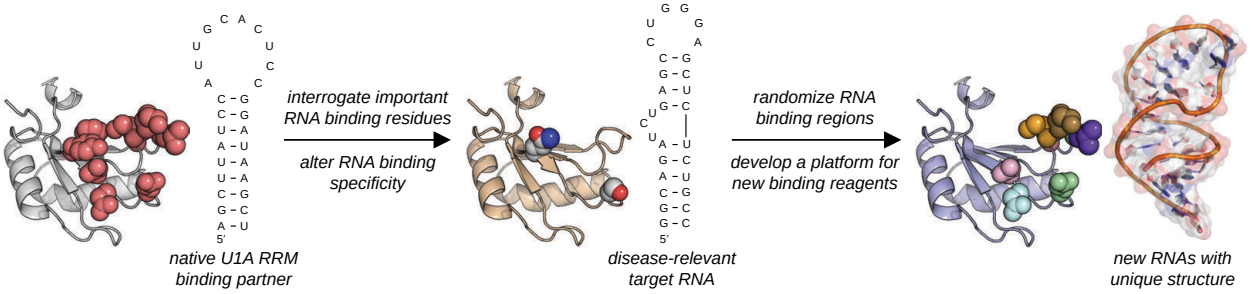


Figure 1.19 Overall strategy for developing a general approach toward new synthetic reagents that recognize and bind physiologically relevant folded RNA molecules. Residues on the RRM U1A known to be important for recognizing it's native RNA binding partner (*left*, highlighted as red spheres) are systematically investigated for their role in recognizing a different RNA target with no sequence homology to the native RNA. Mutations are identified (*center*, shown as spheres) that alter the binding specificity of the RRM scaffold and provide good affinity for the new target, thus showing proof-of-principle that this molecular scaffold can be reprogrammed to bind new RNAs. Select residues on the protein scaffold are randomized to create a combinatorial library of synthetic RRMs. High-throughput screening is then used to identify members that tightly and specifically bind a desired folded RNA target. (PDB codes: 1URN and 2LDL)

reagents to "capture" low energy conformations or tertiary structures, which commonly dictate the function of physiological RNAs.

The RNA-binding mechanism of the U1A RRM has been well characterized through previous biophysical and mutational studies[198-200]. These studies demonstrate that the U1A scaffold is amendable to surface mutations without compromising its structural stability. As stated previously, this is an essential property for protein engineering efforts. Additionally, this wealth of binding information allows us to thoughtfully direct new mutations to putative RNA binding regions on the protein scaffold[201,202] **Figure 1.19** *red spheres*, to alter its RNA binding selectivity. Collectively this makes the U1A scaffold an attractive framework for expanding upon the natural set of RNA binding proteins by serving as a template to engineer or screen for new synthetic RNA binding reagents.

Toward this goal we systematically evaluate U1A's ability to accommodate new RNA

targets through structure activity relationship studies. This is followed by a high-throughput screening effort to identify new RNA-binding mutants by yeast display. To our knowledge, this is the first application of yeast display to screen for new RNA ligands. Our results also represent the first example of a synthetic RRM redesigned to target a disease relevant RNA. Preliminary data indicates the new synthetic proteins bind with mid- to low-nanomolar affinity.

Taken together, the work reported in this thesis represents improved or alternative solutions to the molecular recognition of difficult biopolymers that traditionally evade conventional binding reagents. Inherent to their design, these new approaches should be applicable for targeting a broader range of protein or RNA targets that are of high interest to the general scientific community. New synthetic proteins reported within, or derived from the methods detailed here, will likely find use as basic research tools to interrogate biological systems or as promising therapeutic drug leads.

REFERENCES

(1) Llarrull, L. I.; Testero, S. A.; Fisher, J. F.; Mobashery, S. *Curr Opin Microbiol* **2010**, *13*, 551.

(2) Kovarik, J. M.; Burtin, P. *Expert Opin Emerging Drugs* **2003**, *8*, 47.

(3) Risinger, A. L.; Li, J.; Bennett, M. J.; Rohena, C. C.; Peng, J.; Schriemer, D. C.; Mooberry, S. L. *Cancer Res* **2013**, *73*, 6780.

(4) Madiraju, C.; Edler, M. C.; Hamel, E.; Raccor, B. S.; Balachandran, R.; Zhu, G.; Giuliano, K. A.; Vogt, A.; Shin, Y.; Fournier, J.-H.; Fukui, Y.; Brückner, A. M.; Curran, D. P.; Day, B. W. *Biochemistry* **2005**, *44*, 15053.

(5) Zimmer, S. M.; Liu, J.; Clayton, J. L.; Stephens, D. S.; Snyder, J. P. *J Biol Chem* **2008**, *283*, 27916.

(6) Overington, J. P.; Al-Lazikani, B.; Hopkins, A. L. *Nat Rev Drug Discov* **2006**, *5*, 993.

(7) Rask-Andersen, M.; Almén, M. S.; Schiöth, H. B. *Nat Rev Drug Discov* **2011**, *10*, 579.

(8) Dandapani, S.; Marcaurelle, L. A. *Nat Chem Biol* **2010**, *6*, 861.

(9) Verdine, G. L.; Walensky, L. D. *Clinical Cancer Res* **2007**, *13*, 7264.

(10) Visintin, M.; Melchionna, T.; Cannistraci, I.; Cattaneo, A. *J Biotechnol* **2008**, *135*, 1.

(11) Fuller, J. C.; Burgoyne, N. J.; Jackson, R. M. *Drug Discov Today* **2009**, *14*, 155.

(12) Zinzalla, G.; Thurston, D. E. *Future Medicinal Chemistry* **2009**, *1*, 65.

(13) Sheinerman, F. B.; Norel, R.; Honig, B. *Curr Opin Struc Biol* **1998**, *10*, 153.

(14) Lawrence, M. C.; Colman, P. M. *J Mol Biol* **1993**, *234*, 946.

(15) Jones, S.; Thornton, J. M. *J Mol Biol* **1997**, *272*, 121.

(16) Conte, Lo, L.; Chothia, C.; Janin, J. *J Mol Biol* **1997**, *285*, 2177.

(17) Gallego, J.; Varani, G. *Acc Chem Res* **2001**, *34*, 836.

(18) Luedtke, N. W.; Liu, Q.; Tor, Y. *Biochemistry* **2003**, *42*, 11391.

(19) Tor, Y. *ChemBioChem* **2003**, *4*, 998.

(20) Manning, G. *Science* **2002**, *298*, 1912.

(21) Stelzl, U.; Worm, U.; Lalowski, M.; Haenig, C.; Brembeck, F. H.; Goehler, H.; Stroedicke, M.; Zenkner, M.; Schoenherr, A.; Koeppen, S.; Timm, J.; Mintzlaff, S.; Abraham, C.; Bock, N.; Kietzmann, S.; Goedde, A.; Toksöz, E.; Droege, A.; Krobitsch, S.; Korn, B.; Birchmeier, W.; Lehrach, H.; Wanker, E. E. *Cell* **2005**, *122*, 957.

(22) Pawson, T.; Nash, P. *Gene Dev* **1998**.

(23) Cohen, F. J. *Discov Med* **2003**, *5*, 153.

(24) Bennani, Y. L. *Drug Discov Today* **2012**, *17*, S31.

(25) Strohl, W. R.; Knight, D. M. *Curr Opin Biotechnology* **2009**, *20*, 668.

(26) Berggren, R.; Møller, M.; Moss, R.; Poda, P.; Smietana, K. *Nat Rev Drug Discov* **2012**, *11*, 435.

(27) Hestekin, C. N.; Barron, A. E. *Electrophoresis* **2006**, *27*, 3805.

(28) Lequin, R. M. *Clin Chem* **2005**, *51*, 2415.

(29) Uprichard, S. L.; Boyd, B.; Althage, A.; Chisari, F. V. *P Natl Acad Sci USA* **2005**, *102*, 773.

(30) Pierce, M. *Methods* **1999**, *19*, 213.

(31) Keller, S.; Vargas, C.; Zhao, H.; Piszczek, G.; Brautigam, C. A.; Schuck, P. *Anal. Chem.* **2012**, *84*, 5066.

(32) Mrksich, M.; Sigal, G. B.; Whitesides, G. M. *Langmuir* **1995**, *11*, 4383.

(33) Abbas, A.; Linman, M. J.; Cheng, Q. *Biosens Bioelectron* **2011**, *26*, 1815.

(34) Ciruela, F. *Curr Opin Biotechnology* **2008**, *19*, 338.

(35) Giepmans, B. N. G. *Science* **2006**, *312*, 217.

(36) Choi, J.-W.; Kang, D.-K.; Park, H.; deMello, A. J.; Chang, S.-I. *Anal. Chem.* **2012**, *84*, 3849.

(37) Jameson, D. M.; Ross, J. A. *Chem. Rev.* **2010**, *110*, 2685.

(38) Pfleger, K. D. G.; Eidne, K. A. *Nat Meth* **2006**, *3*, 165.

(39) Song, Y.; Madahar, V.; Liao, J. *Ann Biomed Eng* **2010**, *39*, 1224.

(40) SKLAR, L.; CARTER, M.; EDWARDS, B. *Curr Opin Pharmacology* **2007**, *7*, 527.

(41) Reetz, M. T.; Carballeira, J. D. *Nat Protoc* **2007**, *2*, 891.

(42) Jäckel, C.; Kast, P.; Hilvert, D. *Annu. Rev. Biophys.* **2008**, *37*, 153.

(43) Hoogenboom, H. R. *Nat. Biotechnol.* **2005**, *23*, 1105.

(44) Farinas, E. T.; Bulter, T.; Arnold, F. H. *Curr Opin Biotechnology* **2001**, *12*, 545.

(45) Eijsink, V. G. H.; Gåseidnes, S.; Borchert, T. V.; van den Burg, B. *Biomol Eng* **2005**, *22*, 21.

(46) Bornscheuer, U. T.; Pohl, M. *Curr Opin Chemical Biology* **2001**, *5*, 137.

(47) Lipovsek, D.; Plückthun, A. *J Immunol Methods* **2004**, *290*, 51.

(48) Gai, S. A.; Wittrup, K. D. *Curr Opin Structural Biology* **2007**, *17*, 467.

(49) Daugherty, P. S. *Curr Opin Structural Biology* **2007**, *17*, 474.

(50) Kehoe, J. W.; Kay, B. K. *Chem. Rev.* **2005**, *105*, 4056.

(51) Bradbury, A. R. M.; Sidhu, S.; Dübel, S.; McCafferty, J. *Nat. Biotechnol.* **2011**, *29*, 245.

(52) Löfblom, J. *Biotechnol* **2011**.

(53) Park, H. Y.; Kim, J.; Cho, J. H.; Moon, J. Y.; Lee, S. J.; Yoon, M. Y. *J Biomol Screening* **2011**, *16*, 82.

(54) Horisawa, K. *Nucleic Acids Res.* **2004**, *32*, e169.

(55) Pierzchalski, A.; Mittag, A.; Bocsi, J.; Tarnok, A. *PLoS ONE* **2013**, *8*, e74745.

(56) Valdivia, R. H. *Science* **1997**, *277*, 2007.

(57) Herzenberg, L. A.; Parks, D.; Sahaf, B.; Perez, O.; Roederer, M.; Herzenberg, L. A. *Clin Chem* **2000**, *48*, 1819.

(58) Lönneborg, R.; Varga, E.; Brzezinski, P. *PLoS ONE* **2012**, *7*, e29994.

(59) Macey, M. G. Humana Press, 2005.

(60) van Bloois, E.; Winter, R. T.; Kolmar, H.; Fraaije, M. W. *Trends Biotechnol* **2011**, *29*, 79.

(61) Boder, E. T.; Wittrup, K. D. *Nat. Biotechnol.* **1997**, *15*, 553.

(62) Pepper, L. R.; Cho, Y. K.; Boder, E. T.; Shusta, E. V. *Comb. Chem. High Throughput Screen.* **2006**, *11*, 127.

(63) Fields, S.; Uetz, P.; Giot, L.; Cagney, G.; Mansfield, T. A.; Judson, R. S.; Knight, J. R.; Lockshon, D.; Narayan, V.; Srinivasan, M.; Pochart, P.; Qureshi-Emili, A.; Li, Y.; Godwin, B.; Conover, D.; Kalbfleisch, T.; Vijayadamodar, G.; Yang, M.; Johnston, M.; Rothberg, J. M. *Nature* **2000**, *403*, 623.

(64) Cagney, G.; Uetz, P.; Fields, S. *Meth. Enzymol.* **1998**, *328*, 3.

(65) Yang, M.; Wu, Z.; Fields, S. *Nucleic Acids Res.* **1993**, *23*, 1152.

(66) Wodak, S. J.; Pu, S.; Vlasblom, J.; Seraphin, B. *Mol Cell Proteomics* **2009**, *8*, 3.

(67) Brettner, L. M.; Masel, J. *BMC Syst Biol* **2012**, *6*, 128.

(68) Stynen, B.; Tournu, H.; Tavernier, J.; Van Dijck, P. *Microbiol Mol Biol R* **2012**, *76*, 331.

(69) Remy, I.; Michnick, S. W. *BioTechniques* **2005**, *42*, 137.

(70) Ferrara, F.; Listwan, P.; Waldo, G. S.; Bradbury, A. R. M. *PLoS ONE* **2011**, *6*, e25727.

(71) Michnick, S. W.; Ear, P. H.; Landry, C.; Malleshaiah, M. K.; Messier, V. In *Antibody Engineering*; Methods in Molecular Biology; Humana Press: Totowa, NJ, 2011; Vol. 756, pp. 395–425.

(72) Shekhawat, S. S.; Ghosh, I. *Curr Opin Chemical Biology* **2011**, *15*, 789.

(73) Michnick, S. W.; Ear, P. H.; Manderson, E. N.; Remy, I.; Stefan, E. *Nat Rev Drug Discov* **2007**, *6*, 569.

(74) Galarneau, A.; Primeau, M.; Trudeau, L.-E.; Michnick, S. W. *Nat. Biotechnol.* **2002**, *20*, 619.

(75) Hashimoto, T.; Adams, K. W.; Fan, Z.; McLean, P. J.; Hyman, B. T. *J Biol Chem* **2011**, *286*, 27081.

(76) Wilson, C. G.; Magliery, T. J.; Regan, L. *Nat Meth* **2004**, *1*, 255.

(77) Sarkar, M.; Magliery, T. J. *Mol. BioSyst.* **2008**, *4*, 599.

(78) Cabantous, S.; Waldo, G. S. *Nat Meth* **2006**, *3*, 845.

(79) Remy, I.; Michnick, S. W. *P Natl Acad Sci USA* **1998**, *96*, 5394.

(80) Tarassov, K.; Messier, V.; Landry, C. R.; Radinovic, S.; Molina, M. M. S.; Shames, I.; Malitskaya, Y.; Vogel, J.; Bussey, H.; Michnick, S. W. *Science* **2008**, *320*, 1465.

(81) Jackrel, M. E.; Cortajarena, A. L.; Liu, T. Y.; Regan, L. *ACS Chem. Biol.* **2010**, *5*, 553.

(82) Magliery, T. J.; Wilson, C. G. M.; Pan, W.; Mishler, D.; Ghosh, I.; Hamilton, A. D.; Regan, L. *J. Am. Chem. Soc.* **2005**, *127*, 146.

(83) Magliery, T. J.; Regan, L. *Methods Biochem Anal* **2005**, *47*, 391.

(84) Lawrence, M. S.; Phillips, K. J.; Liu, D. R. *J. Am. Chem. Soc.* **2007**, *129*, 10110.

(85) Blakeley, B. D.; Chapman, A. M.; McNaughton, B. R. *Mol. BioSyst.* **2012**, *8*, 2036.

(86) Sheng, J.; Gan, J.; Huang, Z. *Med. Res. Rev.* **2013**, *33*, 1119.

(87) Vester, B.; Wengel, J. *Biochemistry* **2004**, *43*, 13233.

(88) Havens, M. A.; Duelli, D. M.; Hastings, M. L. *WIREs RNA* **2013**, *4*, 247.

(89) Guan, L.; Disney, M. D. *ACS Chem. Biol.* **2012**, *7*, 73.

(90) Bennett, C. F.; Swayze, E. E. *Annu. Rev. Pharmacol. Toxicol.* **2010**, *50*, 259.

(91) Extance, A. *Nat Rev Drug Discov* **2009**, *8*, 917.

(92) Foloppe, N.; Matassova, N.; Aboul-ela, F. *Drug Discov Today* **2006**, *11*, 1019.

(93) Leaman, D. W. *Expert Opin. Drug Discov.* **2008**, *3*, 997.

(94) Zaman, G. *Drug Discov Today* **2003**, *8*, 297.

(95) Ling, H.; Fabbri, M.; Calin, G. A. *Nat Rev Drug Discov* **2013**.

(96) Cho, J.; Parks, M. E.; Dervan, P. B. *P Natl Acad Sci USA* **1994**, *92*, 10389.

(97) Dervan, P. B.; White, S.; Szewczyk, J. W.; Turner, J. M.; Baird, E. E. *Nature* **1998**, *391*, 468.

(98) Dervan, P. B.; Edelson, B. S. *Curr Opin Structural Biology* **2003**, *13*, 284.

(99) Raskatov, J. A.; Puckett, J. W.; Dervan, P. B. *Bioorgan Med Chem* **2014**.

(100) Kopka, M. L.; Yoon, C.; Goodsell, D.; Pjura, P.; Dickerson, R. E. *P Natl Acad Sci USA* **1985**, *82*, 1376.

(101) Wurtz, N. R.; Turner, J. M.; Baird, E. E.; Dervan, P. B. *Org Lett* **2001**.

(102) Nickols, N. G.; Szablowski, J. O.; Hargrove, A. E.; Li, B. C. *Mol Cancer Ther* **2013**.

(103) Yang, F.; Nickols, N. G.; Li, B. C.; Marinov, G. K.; Said, J. W.; Dervan, P. B. *P Natl Acad Sci USA* **2013**, *110*, 1863.

(104) Kang, J.; Meier, J. L.; Dervan, P. B. *J Am Chem Soc* **2014**.

(105) Cruz, J. A.; Westhof, E. *Cell* **2009**, *136*, 604.

(106) Bailor, M. H.; Sun, X.; Al-Hashimi, H. M. *Science* **2010**, *327*, 202.

(107) Dethoff, E. A.; Chugh, J.; Mustoe, A. M.; Al-Hashimi, H. M. *Nature* **2012**, *482*, 322.

(108) Deigan, K. E.; Li, T. W.; Mathews, D. H.; Weeks, K. M. *P Natl Acad Sci U.S.A.* **2009**, *106*, 97.

(109) Lipinski, C. A. *Drug Discov Today: Technologies* **2004**, *1*, 337.

(110) Hermann, T. *Biopolymers* **2003**, *70*, 4.

(111) Thomas, J. R.; Hergenrother, P. J. *Chem Rev* **2008**, *108*, 1171.

(112) Gohlke, H.; Klebe, G. *Angew Chem Int Ed Engl* **2002**, *41*, 2644.

(113) Williamson, J. R. *Nat Struct Biol* **2000**, *7*, 834.

(114) Nutiu, R.; Li, Y. *J Am Chem Soc* **2003**, *125*, 4771.

(115) Blackburn, E. H. *Cell* **2001**, *106*, 661.

(116) Jones, S.; Daley, D. T.; Luscombe, N. M.; Berman, H. M.; Thornton, J. M. *Nucleic Acids Res.* **2001**, *29*, 943.

(117) Lunde, B. M.; Moore, C.; Varani, G. *Nat Rev Mol Cell Biol* **2007**.

(118) Evans, B. E.; Rittle, K. E.; Bock, M. G. *J Med Chem* **1988**, *31*, 2235.

(119) Welsch, M. E.; Snyder, S. A.; Stockwell, B. R. *Curr Opin Chemical Biology* **2010**, *14*, 347.

(120) Koehn, F. E.; Carter, G. T. *Nat Rev Drug Discov* **2005**, *4*, 206.

(121) Wells, J. A.; McClendon, C. L. *Nature* **2007**, *450*, 1001.

(122) Barelier, S.; Krimm, I. *Curr Opin Chemical Biology* **2011**, *15*, 469.

(123) Petrone, P. M.; Wassermann, A. M.; Lounkine, E.; Kutchukian, P.; Simms, B.; Jenkins, J.; Selzer, P.; Glick, M. *Drug Discov Today* **2013**, *18*, 674.

(124) Walter, F.; Vicens, Q.; Westhof, E. *Curr Opin Chemical Biology* **1999**, *3*, 694.

(125) Vicens, Q.; Westhof, E. *ChemBioChem* **2003**, *4*, 1018.

(126) Chittapragada, M.; Roberts, S.; Ham, Y. W. *Perspect Medicin Chem* **2009**, *3*, 21.

(127) Waksman, S. A.; Lechevalier, H. A. *Science* **1949**, *109*, 305.

(128) Kohanski, M. A.; Dwyer, D. J.; Collins, J. J. *Nat Rev Micro* **2010**, *8*, 423.

(129) Miller, G. H.; Arcieri, G.; Weinstein, M. J. *Antimicrob Agents Ch* **1976**.

(130) Hainrichson, M.; Nudelman, I.; Baasov, T. *Org Biomol Chem* **2008**.

(131) Berger, A.; Kretzer, V.; Gludovatz, P.; Heinze, G.; Haiden, N.; Pollak, A. *Acta Paediatrica* **2007**, *93*, 356.

(132) Blatter, M.; Fluckiger, U.; Entenza, J.; Glauser, M. P.; Francioli, P. *Antimicrob Agents Ch* **1993**, *37*, 1971.

(133) Solera, J.; Espinosa, A.; Geijo, P.; Martínez-Alfaro, E.; Sáez, L.; Sepúlveda, M. A.; Ruiz-Ribó, M. D. *Clin. Infect. Dis.* **1996**, *22*, 441.

(134) Ryu, D. H.; Rando, R. R. *Bioorg Med Chem* **2001**, *9*, 2601.

(135) Liang, F.-S.; Greenberg, W. A.; Hammond, J. A.; Hoffmann, J.; Head, S. R.; Wong, C.-H. *P Natl Acad Sci USA* **2006**, *103*, 12311.

(136) Rando, R. R. *Specificity in the Binding of Aminoglycosides to RNA, RNA-Binding Antibiotics* **2001**.

(137) Liu, X.; Thomas, J. R.; Hergenrother, P. J. *J. Am. Chem. Soc.* **2004**, *126*, 9196.

(138) Thomas, J. R.; Liu, X.; Hergenrother, P. J. *J. Am. Chem. Soc.* **2005**, *127*, 12434.

(139) Thomas, J. R.; Liu, X.; Hergenrother, P. J. *Biochemistry* **2006**, *45*, 10928.

(140) Disney, M. D.; Barrett, O. J. *Biochemistry* **2007**, *46*, 11223.

(141) Paul, D. J.; Seedhouse, S. J.; Disney, M. D. *Nucleic Acids Res.* **2009**, *37*, 5894.

(142) Disney, M. D.; Labuda, L. P.; Paul, D. J.; Poplawski, S. G.; Pushechnikov, A.; Tran, T.; Velagapudi, S. P.; Wu, M.; Childs-Disney, J. L. *J. Am. Chem. Soc.* **2008**, *130*, 11185.

(143) Pushechnikov, A.; Lee, M. M.; Childs-Disney, J. L.; Sobczak, K.; French, J. M.; Thornton, C. A.; Disney, M. D. *J. Am. Chem. Soc.* **2009**, *131*, 9767.

(144) Parkesh, R.; Childs-Disney, J. L.; Nakamori, M.; Kumar, A.; Wang, E.; Wang, T.; Hoskins, J.; Tran, T.; Housman, D.; Thornton, C. A.; Disney, M. D. *J. Am. Chem. Soc.* **2012**, *134*, 4731.

(145) Childs-Disney, J. L.; Yildirim, I.; Park, H.; Lohman, J. R.; Guan, L.; Tran, T.; Sarkar, P.; Schatz, G. C.; Disney, M. D. *ACS Chem. Biol.* **2014**, *9*, 538.

(146) Hancock, R. E.; Chapple, D. S. *Antimicrob Agents Ch* **1999**, *43*, 1317.

(147) Fernandez-Lopez, S.; Kim, H. S.; Choi, E. C.; Delgado, M.; Granja, J. R.; Khasanov, A.; Kraehenbuehl, K.; Long, G.; Weinberger, D. A.; Wilcoxen, K. M.; Ghadiri, M. R. *Nature* **2001**, *412*, 452.

(148) Lalonde, M. S.; Lobritz, M. A.; Ratcliff, A.; Chamanian, M.; Athanassiou, Z.; Tyagi, M.; Wong, J.; Robinson, J. A.; Karn, J.; Varani, G.; Arts, E. J. *PLoS Pathog* **2011**, *7*, e1002038.

(149) Craik, D. J.; Fairlie, D. P.; Liras, S.; Price, D. *Chem Biol Drug Design* **2012**, *81*, 136.

(150) Wilson, D. N. *Nature* **2014**, *12*, 35.

(151) Blaha, G. M.; Polikanov, Y. S.; Steitz, T. A. *Curr Opin Structural Biology* **2012**, *22*, 750.

(152) Stanley, R. E.; Blaha, G.; Grodzicki, R. L.; Strickler, M. D.; Steitz, T. A. *Nature* **2010**, *17*, 289.

(153) Akbergenov, R.; Shcherbakov, D.; Matt, T.; Duscha, S.; Meyer, M.; Wilson, D. N.; Bottger, E. C. *Antimicrob Agents Ch* **2011**, *55*, 4712.

(154) Burns, V. A.; Bobay, B. G.; Basso, A.; Cavanagh, J.; Melander, C. *Bioorg Med Chem Lett* **2008**, *18*, 565.

(155) Baek, S.; Kutchukian, P. S.; Verdine, G. L.; Huber, R.; Holak, T. A.; Lee, K. W.;

Popowicz, G. M. *J. Am. Chem. Soc.* **2012**, *134*, 103.

(156) Henchey, L. K.; Jochim, A. L.; Arora, P. S. *Curr Opin Chem Biol* **2008**, *12*, 692.

(157) Moellering, R. E.; Cornejo, M.; Davis, T. N.; Del Bianco, C.; Aster, J. C.; Blacklow, S. C.; Kung, A. L.; Gilliland, D. G.; Verdine, G. L.; Bradner, J. E. *Nature* **2009**, *462*, 182.

(158) Walensky, L. D.; Kung, A. L.; Escher, I.; Malia, T. J.; Barbuto, S. *Science* **2004**.

(159) Athanassiou, Z.; Dias, R. L. A.; Moehle, K.; Dobson, N.; Varani, G.; Robinson, J. A. *J. Am. Chem. Soc.* **2004**, *126*, 6906.

(160) Davidson, A.; Leeper, T. C.; Athanassiou, Z.; Patora-Komisarska, K.; Karn, J.; Robinson, J. A.; Varani, G. *P Natl Acad Sci USA* **2009**, *106*, 11931.

(161) Krishnamurthy, M.; Simon, K.; Orendt, A. M.; Beal, P. A. *Angew. Chem.* **2007**, *119*, 7174.

(162) Leeper, T. C.; Athanassiou, Z.; Dias, R. L. A.; Robinson, J. A.; Varani, G. *Biochem* **2005**, *44*, 12362.

(163) Bardaro, M. F.; Shajani, Z.; Patora-Komisarska, K.; Robinson, J. A.; Varani, G. *Nucleic Acids Res.* **2009**, *37*, 1529.

(164) Davidson, A.; Patora-Komisarska, K.; Robinson, J. A.; Varani, G. *Nucleic Acids Res.* **2011**, *39*, 248.

(165) Manna, A. K.; Kumar, A.; Ray, U.; Das, S.; Basu, G.; Roy, S. *Antiviral Res* **2013**, *97*, 223.

(166) Ali, N.; Pruijn, G.; Kenan, D. J.; Keene, J. D. *J Biol Chem* **2000**, *275*, 27531.

(167) Mondal, T.; Ray, U.; Manna, A. K.; Gupta, R.; Roy, S.; Das, S. *J Virol* **2008**, *82*, 11927.

(168) Egholm, M.; Buchardt, O.; Nielsen, P. E. *J Am Chem Soc* **1992**, *114*, 1895.

(169) Nielsen, P. E. *Curr Opin Structural Biology* **1999**, *9*, 353.

(170) Nielsen, E. *Handb Exp Pharmacol* **2006**, 395.

(171) Boutimah-Hamoudi, F.; Leforestier, E.; Senamaud-Beaufort, C.; Nielsen, P. E.; Giovannangeli, C.; Saison-Behmoaras, T. E. *Nucleic Acids Res.* **2007**, *35*, 3907.

(172) Nielsen, P. E. *Chem. Biodivers.* **2010**, *7*, 786.

(173) Yuan, L.; Kurek, I.; English, J.; Keenan, R. *Microbiol Mol Biol Rev* **2005**, *69*, 373.

(174) Lutz, S. *Curr Opin Biotechnol* **2010**, *21*, 734.

(175) Brustad, E. M.; Arnold, F. H. *Curr Opin Chemical Biology* **2011**, *15*, 201.

(176) Esvelt, K. M.; Carlson, J. C.; Liu, D. R. *Nature* **2012**, *472*, 499.

(177) Hoebenreich, S.; Zilly, F. E.; Rocha, C. A. *ACS Syn Biol* **2014**, 140612140825003.

(178) Neylon, C. *Nucleic Acids Res.* **2004**, *32*, 1448.

(179) Cobb, R. E.; Si, T.; Zhao, H. *Curr Opin Chemical Biology* **2012**, *16*, 285.

(180) Zamore, P. D.; Williamson, J. R.; Lehmann, R. *RNA* **1997**, *3*, 1421.

(181) Wharton, R. P.; Sonoda, J.; Lee, T.; Patterson, M.; Murata, Y. *Mol. Cell* **1998**.

(182) Wang, X.; Zamore, P. D.; Hall, T. M. *Mol. Cell* **2001**, *7*, 855.

(183) Wang, X.; McLachlan, J.; Zamore, P. D.; Hall, T. *Cell* **2002**.

(184) Cheong, C.-G.; Hall, T. M. T. *P Natl Acad Sci USA* **2006**, *103*, 13635.

(185) Wang, Y.; Wang, Z.; Tanaka Hall, T. M. *FEBS J* **2013**, *280*, 3755.

(186) Barkan, A.; Rojas, M.; Fujii, S.; Yap, A.; Chong, Y. S.; Bond, C. S.; Small, I. *PLoS Genet* **2012**, *8*, e1002910.

(187) Fujii, S.; Small, I. *New Phytologist* **2011**, *191*, 37.

(188) Yagi, Y.; Nakamura, T.; Small, I. *Plant J* **2014**, *78*, 772.

(189) Romero, P. A.; Arnold, F. H. *Nat Rev Mol Cell Biol* **2009**, *10*, 866.

(190) Hietpas, R. T.; Jensen, J. D.; Bolon, D. N. A. *P Natl Acad Sci USA* **2011**, *108*, 7896.

(191) Romero, P. A.; Krause, A.; Arnold, F. H. *P Natl Acad Sci USA* **2013** *110*, E193

(192) Yoshikuni, Y.; Ferrin, T. E.; Keasling, J. D. *Nature* **2006**, *440*, 1078.

(193) Nulf, C. J.; Corey, D. *Nucleic Acids Res.* **2004**.

(194) Crick, F.; Watson, J. *Nature* **1953**.

(195) Maris, C.; Dominguez, C.; Allain, F. H. T. *FEBS J* **2005**, *272*, 2118.

(196) Bejerano, G. *Science* **2004**, *304*, 1321.

(197) Cléry, A.; Blatter, M.; Allain, F. *Curr Opin Structural Biology* **2008**.

(198) Varani, G.; Nagai, K. *Annu Rev Biophys Biomol Struct* **1998**, *27*, 407.

(199) Jacks, A.; Babon, J.; Kelly, G.; Manolaridis, I.; Cary, P. D.; Curry, S.; Conte, M. R. *Structure* **2003**, *11*, 833.

(200) Kranz, J. K.; Hall, K. B. *J Mol Biol* **1998**, *275*, 465.

(201) Auweter, S. D.; Oberstrass, F. C.; Allain, F. H. T. *Nucleic Acids Res.* **2006**, *34*, 4943.

(202) Bentley, R. C.; Keene, J. D. *Mol Cell Biol* **1991** *11*, 1829.

CHAPTER TWO

SPLIT-SUPERPOSITIVE GREEN FLUORESCENT PROTEIN REASSEMBLY IS A FAST, EFFICIENT AND ROBUST METHOD FOR DETECTING PROTEIN-PROTEIN INTERACTIONS IN LIVE CELLS

2.1 Introduction

Protein-Protein Interactions (PPI's) dictate a multitude of cellular events, including disease-relevant cellular function and fate. Identifying disease-relevant PPI's in the context of complex environments, such as a living cell, remains a central challenge to the field of proteomics. Moreover, the identification of synthetic proteins that bind disease-relevant proteins, and inhibit their physiological interactions, is needed to identify new proteinaceous drug leads and basic research tools.

Common methods for identifying PPIs in living cells include yeast 2-hybrid[1,2] and protein complementation assays[3,4] **Figure 2.1**. Yeast 2-hybrid assays are relatively limited as a result of two primary factors. First, false positives routinely contaminate binding results[2]. Second, yeast have relatively short doubling times (~2 hours) and limited transfection efficiency (~10^7 transformants)[5]. Conversely, protein complementation assays (PCAs) can be performed in *E.coli*, which enable higher transformation efficiencies (~10^{10}) and faster doubling time (~20 minutes)[6]. Taken together, PCAs in *E.coli* enable more PPIs to be probed in a shorter amount of time.

To implement a PCA, potential interacting proteins are each tethered to non-functional

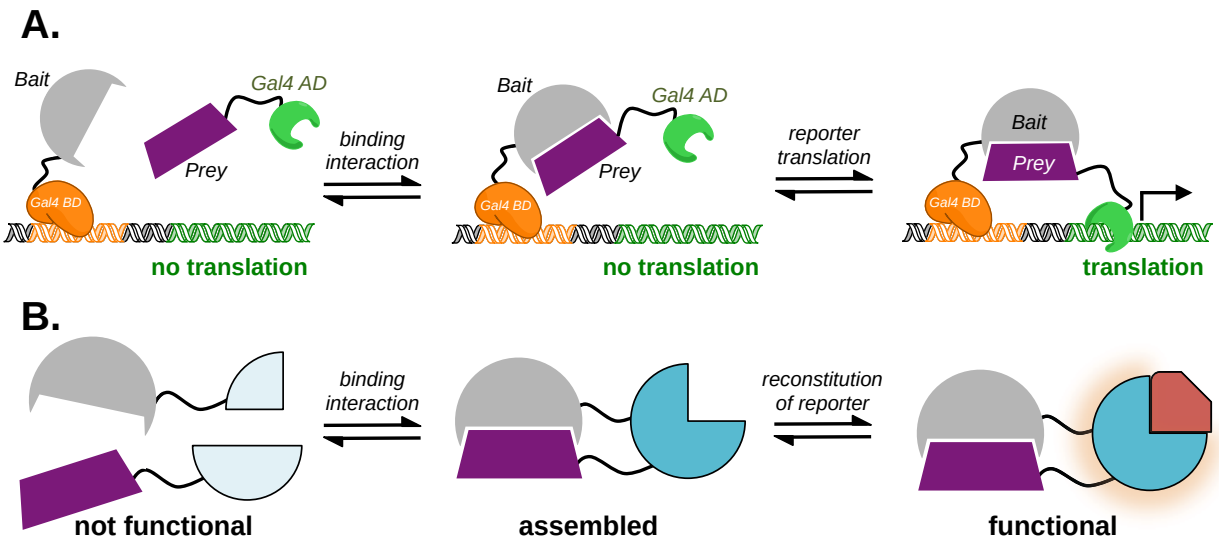


Figure 2.1 A Shown is a cartoon schematic of a yeast hybrid assay. The bait protein (*grey*) is expressed as a fusion with the Gal4 binding domain (BD) (*orange*). The Gal4 BD of the chimera protein remains bound to the upstream activation sequence on the DNA. Tight binding between bait and prey (*purple*) proteins induces a high effective molarity of the Gal4 activation domain (AD) (*green*). The Gal4 AD up-regulates translation of a downstream reporter protein to indicate binding. **B** Shown is a cartoon schematic of a protein complementation assay. Two interacting proteins (grey and purple) are each fused to non-functional fragments of a split-reporter protein (light blue). Association of the interacting proteins facilitates the reassembly of the fully functional reporter protein (dark blue). The reporter protein facilitates a phenotypic change in the host organism, shown here as the modification of an exogenous reagent to produce a visible signal.

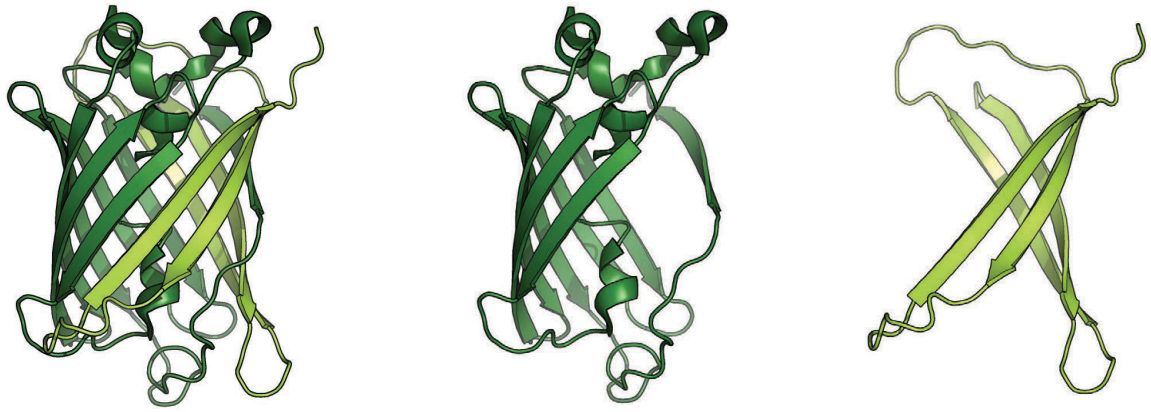
complimentary fragments of a reporter protein and expressed concomitantly inside a cell. Reassembly of the fully functional reporter protein is contingent upon association between the proteins of interest **Figure 2.1B**. The non-functional fragments do not reassemble in the absence of a binding interaction and consequently, no phenotypic change is observed.

A variety of reporter proteins have been adapted for PCAs, including those that generate a fluorescent signal, catalyze a colorimetric or fluorescent reaction, or enable a functional pathway essential to the organism's survival. Specific examples include β-lactamase[7], β-galactosidase[8], dihydrofolate reductase[9], ubiquitin[10], and Green Fluorescent Protein (GFP)[11]. Unlike some other examples, the GFP reporter does not require addition of exogenous dyes or metabolites to

generate an observable signal[11,12]. Additionally, the excitation/emmission spectrum of GFP is generally compatible with most fluorometers. The assay is also incredibly sensitive; measurable dissociation constants are as high as 1 mM with negligible background signal[12].

Still, despite the simplicity of applying split-GFP to the study of PPIs, significant limitations have impeded its widespread use. Reassembly of sufficient functional GFP molecules to evolve an observable fluorescence signal typically requires overnight incubation at room temperature (~25 °C) [12]. This devalues the resultant positive binding indications as behavior of a PPI can vary dramatically between room temperature and physiological temperature. Slow and inefficient reassembly is likely due to aggregation of the split protein fragments. Isolating the two halves exposes the hydrophobic core in the GFP β-barrel structure **Figure 2.2**. Aggregation into inclusion bodies occludes binding partners from each other. This inherently detracts from a large-scale screening effort, decreasing the potential interaction diversity.

Recent work from Professor David Liu's lab at Harvard University has shown that increasing the theoretical net-charge of a protein can bestow superior solubility and stability properties[13]. The resulting "supercharged" protein resists aggregation and in some cases can actually refold after thermal denaturation. These properties are dependent on theoretical net charge as proteins with a similar number of charged residues, but a low theoretical net charge, do not demonstrate the same behavior. This suggests that the high net-charge lowers folding energy and/or restricts access to misfolding pathways[14]. The split fragments of sg100GFP have low net charges, both are -4 **Figure 2.3**. We saw protein supercharging as a compelling opportunity to potentially remediate the known aggregation problems associated with split-GFP.



whole green fluorescent protein (GFP) *N-GFP* *C-GFP*

Figure 2.2 The full GFP molecule (*left*) is in the shape of a barrel. The center of the barrel is largely composed of hydrophobic residues. Upon splitting the protein into two fragments, N-GFP and C-GFP (*center* and *right*), these residues become solvent accessible. This can lead to insolubility and aggregation of the split fragments. (PDB code: 2B3Q)

Supercharging the fragments of split-GFP may endow similar solubility properties, as observed with whole supercharged GFP[15]. Minimizing the percentage of fragments lost to aggregation should increase the fluorescence signal over background and improve the utility of split-GFP for high-throughput screening applications **Figure 2.3**. Moreover, increased split-GFP fragment solubility may permit the use of sparingly soluble proteins or peptides in the resulting PCA.

2.2 Split Superpositive-GFP Gives a Brighter Fluorescent Signal

Toward the development of an improved split-GFP assay, we first looked to establish the baseline efficacy of the original assay. The plasmids encoding split-sg100GFP fragments, which derive from an enhanced stability variant of GFP[11], were kindly provided by Lynn Regan's lab. A high affinity binding pair of leucine zippers, denoted Z and Z′, serve as the initial positive binding control. Each is fused to either the N- or C- fragments of split-sg100GFP, respectively.

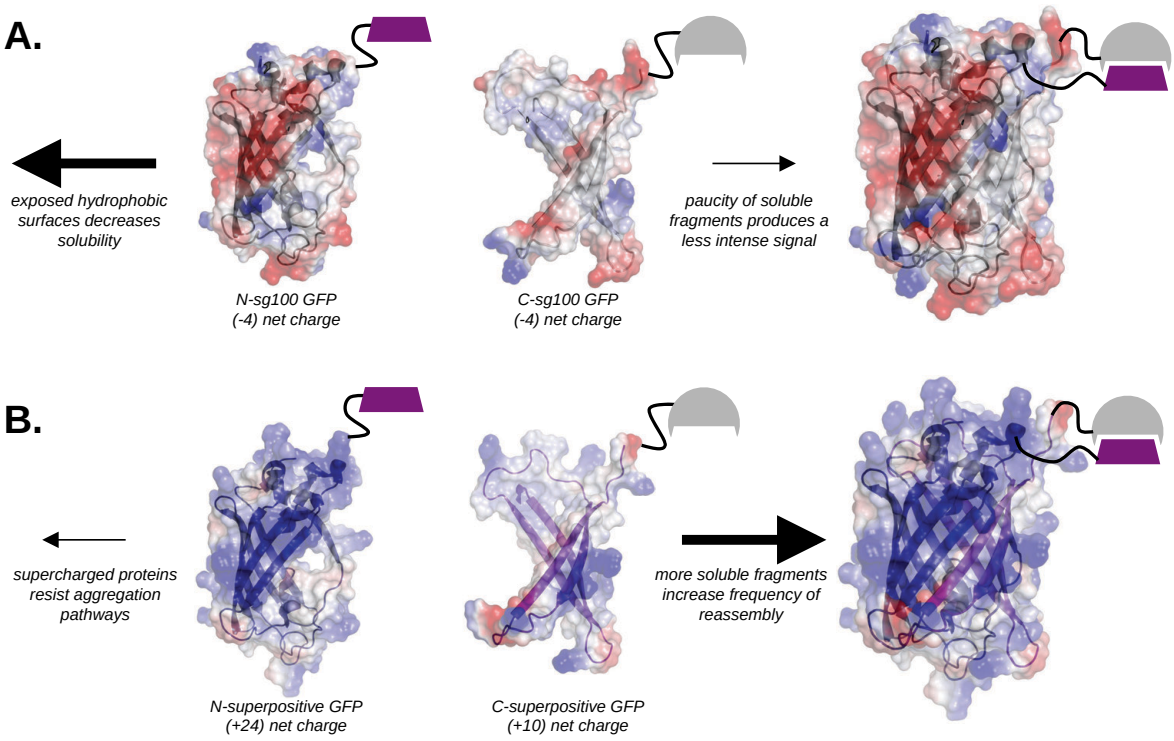


Figure 2.3 A The exposed hydrophobic core of split-sg100GFP causes solubility issues and likely leads to increased aggregation of the fused fragments. **B** The high net positive charge on split-spGFP fragments allows the fusions to resist aggregation pathways and increases solubility. The fragments are shown with vacuum electrostatics generated in pymol to highlight the differences in charge. (PDB code: 2B3Q)

The fusions are each expressed off of separate plasmids, N-sg100GFP-Z-pET11a and Z′-C-sg100GFP-pMRBAD, which harbor unique antibiotic resistance markers. Concomitant expression of both chimeras in BL21 *E.coli* allows the leucine zippers to associate, facilitating reassembly of the split fragments. The fluorescence signal generated from split-sg100 reassembly is shown in **Figure 2.4A**.

With a working positive control in hand, we next looked to evaluate if supercharging the split-GFP fragments could improve the overall performance of the assay. A superpositive variant of GFP, which closely emulates a supercharged-GFP reported by Liu and coworkers[13], was split to match the analogous sequences of the split-sg100 fragments. Split-superpositive GFP

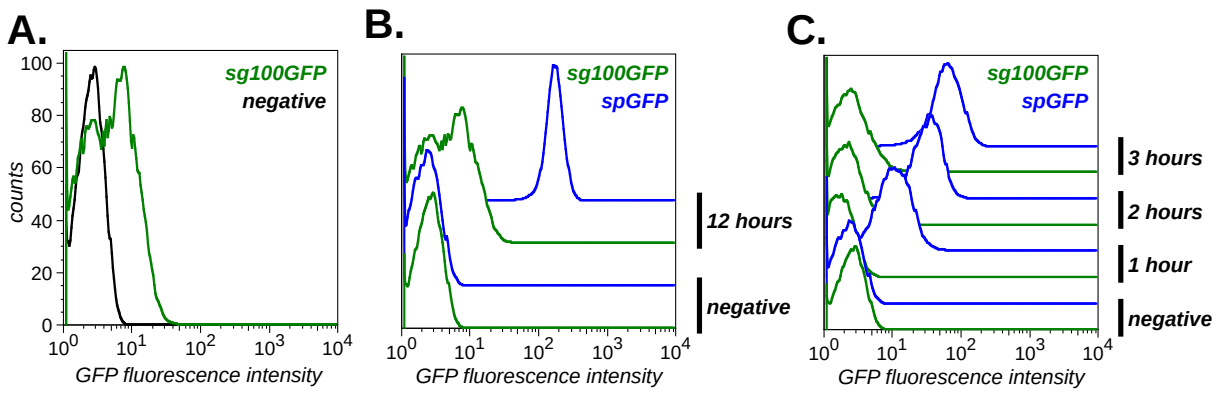


Figure 2.4 A The original split-GFP assay, based on sg100GFP, demonstrates a modest increase in fluorescence as measured by flow cytometry. **B** A side-by-side comparison of split-GFP assays, based on either sg100GFP or superpositive GFP (spGFP). Split-spGFP shows a much brighter and cleaner signal after 12 hours at 25 °C. **C** Split-spGFP evolves a fluorescent signal much faster than split-sg100GFP, suggesting a much more efficient reassembly process.

(spGFP) was cloned into the positive control vectors, replacing split-sg100 fragments, to generate NspGFP-Z-pET11a and Z′-CspGFP-pMRBAD. The net charges of N- and C- split spGFP are +24 and +10, respectively **Figure 2.3B**. The fluorescent intensity from each PCA was quantified using fluorescence flow cytometry. The signal for both assays appeared to saturate after 12 hours. A side-by-side comparison of the two assays showed that split-spGFP generated a 75-fold increase in GFP fluorescence over background, representing a 72.8-fold brighter signal than split-sg100GFP **Figure 2.4B**. This observation suggests that supercharging the split-GFP fragments does indeed circumvent barriers to reassembly, likely by preventing access to aggregation pathways and/or increasing the solubility of the split fragments. The significantly brighter signal enables an easier binary delineation between binding and non-binding protein pairs. Moreover, the increased signal over background can be of great utility for selecting positive colonies from an agar plate in a low-throughput screen and for setting positive thresholds in a high-throughput flow cytometry screen.

2.3 Split-spGFP Reassembly Enables Faster Signal Generation

Split-sg100GFP requires overnight incubation following induction before appreciable fluorescence is observed[16]. In light of the increased fluorescence signal observed, likely due to more efficient reassembly of split-spGFP, we hypothesized that if split-spGFP may also evolve a fluorescence signal much faster than split-sg100GFP. To test this, we arranged a side-by-side comparison of the two systems at varying time points of 1, 2, and 3 hours. The fluorescence signal from split-sg100 GFP showed very little change before three hours. In contrast, split-spGFP demonstrated a time-dependent increase in fluorescence, showing an appreciable signal after only one hour of induction **Figure 2.4C**. This further suggests that, like previously reported supercharged proteins, the supercharged fragments of split-spGFP have improved stability and less protein is lost to aggregation. This inherently increases the efficiency of reassembly as two complementary fragments have a better likelihood of engaging in a productive collision. Using the more efficient split-spGFP assay, protein-protein binding interactions *in vivo* can be evaluated after only one hour, expediting PPI evaluation.

2.4 Split-spGFP Reassembly is More Efficient Than Split-frGFP Reassembly

Magliery and coworkers devised an improved split-GFP assay based on a variant engineered for enhanced folding properties[17]. Split-frGFP represents an improvement over split-sg100GFP, likely resulting from a lowered folding energy. We looked to compare the new split-spGFP assay to the split-frGFP variation using the same leucine zipper binding pair as a control.

In a direct comparison after 12 hours of incubation, split-spGFP showed 22% greater GFP fluorescence intensity over background than split-frGFP by flow cytometry **Figure 2.5A**, suggesting that the superpositive variant allows for more efficient reassembly.

To further compare the efficiency of reassembly, we quantified fluorescence intensity over background at 1, 2, and 3 hour time points for the two assays. While both systems generated a fluorescence signal much quicker than split-sg100GFP, split-spGFP gave a brighter signal at each time point **Figure 2.5B**. These data further suggest that split-spGFP facilitates a more efficient reassembly process, leading to a brighter signal in less time.

2.5 Split-spGFP Reassembly is Brighter at Physiological Temperature

Monitoring protein-protein interactions at physiological temperature is of significant importance for the study of human disease states as well as the pursuit of proteinaceous drug reagents. Interactions observed at lower temperatures, such as those required in the traditional split-sg100GFP assay, may not have biological significance. Toward this end, we evaluated the more efficient reporter systems, split-frGFP and split-spGFP, for leucine zipper-mediated reassembly at 37 °C. Amazingly, both systems showed appreciable fluorescence, while the signal over background from split-spGFP was 22% higher **Figure 2.5C**. Also of note, the population distribution of cells expressing split-spGFP chimeras was more narrow, suggesting more uniform and stable behavior at this elevated temperature. Despite the increased ambient energy in the system, split-spGFP fragments still seems to resist aggregation pathways. Collectively, split-

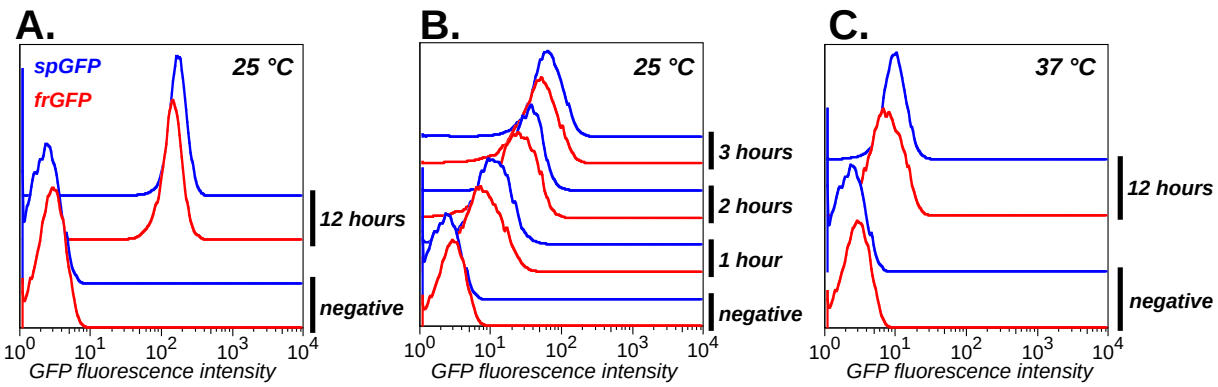


Figure 2.5 A side-by-side comparison of split-spGFP and split-frGFP to monitor a leucine zipper-mediated binding interaction **A** After 12 hours at 25°C, the split-spGFP system generates a 22% brighter signal over background than split-frGFP. **B** Split-spGFP shows faster reassembly than split-frGFP over the course of 3 hours post induction. **C** Split-spGFP also emits a 24% brighter signal than split-frGFP over background after incubation at physiological temperature, 37 °C.

spGFP is a robust method to provide a clean and easily observable indication of protein-protein binding in biologically relevant contexts.

2.6 Reassembly is Not Dependent on Fragment Orientation

Previously reported split-GFP assays have been used to evaluate binding systems involving helical assembly[11,17,18], like leucine zippers. In these one-dimensional complexes, the N- and C-termini align together on the same side of the interaction **Figure 2.6A**. Empirically, this presents an optimal arrangement for the two halves of split-GFP to productively come together and reassemble. However, many protein-protein binding interactions do not maintain such ideal positioning of the N- and C-termini. For broader utility of this assay we looked to evaluate a PPI between two globular proteins.

The PPI binding pair Pdar and Prb bind with a low dissociation constant (180 pM)[19]. The

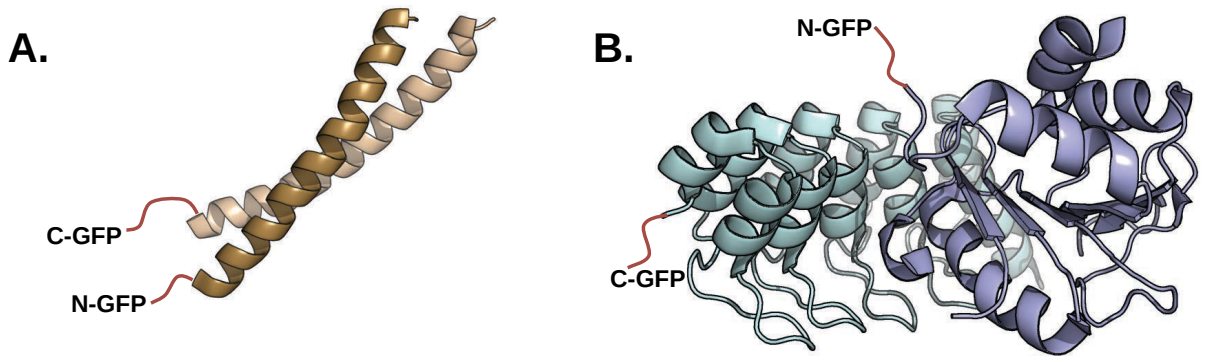


Figure 2.6 A The positive control leucine zipper binding interaction aligns the two fragments of split-GFP into close proximity.**B** The designed proteins Pdar and Prb are more representative of most globular protein-protein interactions. The fragments of split-GFP are fused to termini that are segregated in space, relative to the leucine zipper interaction. Leucine zippers PDB code: 1SER, Pdar and Prb (PDB code: 2DVW)

N- and C-termini of these globular proteins are secluded from one another, relative to the leucine zippers **Figure 2.6B**. This PPI would allow us to evaluate whether split-GFP reassembly is dependent on the positioning of the two split fragments. Pdar and Prb were cloned as fusions with split-frGFP and with split-spGFP fragments. Reassembly of split-frGFP and split-spGFP, mediated by Pdar-Prb, was monitored at 37 °C by flow cytometry. The fluorescence signal saturated after 12 hours in both systems. At this time point, the Pdar/Prb split-spGFP fusions showed a 27% more intense fluorescence signal over background than the equivalent split-frGFP fusions **Figure 2.7A**. This result further suggests that split-spGFP reassembles more efficiently than split-frGFP. Additionally, this demonstrates that split-spGFP is a robust and general method to detect binding interactions of globular proteins *in vivo*. Contrary to previous convention[20], we posit that functional reassembly results from high effective molarity of the complementary fragments, rather than a productive collision that must be guided by the fused interacting proteins.

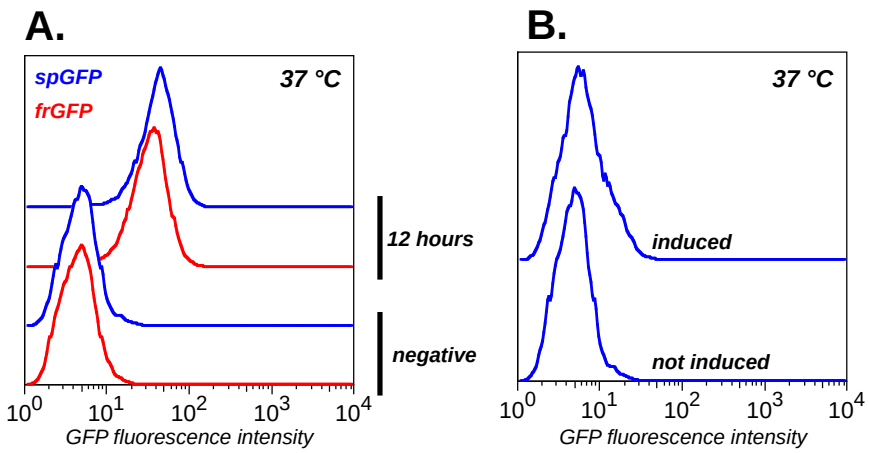


Figure 2.7 A A side-by-side comparison of split-spGFP and split-frGFP to measure binding between the globular proteins Prb and Pdar. Split-spGFP reassembly shows a 27% brighter signal over background. **B** Negative control of non-interacting pairs.

Of significant concern was that the supercharged fragments had been stabilized so dramatically that they may spontaneously reassemble, even in the absence of an interacting protein pair. To evaluate this proposition, we tested a pair of non-binidng proteins in the split-spGFP system. Gankyrin is a stable globular protein that shares a tertiary ankyrin repeat structure with Pdar[21]. However, Gankyrin has a dramatically different primary structure and therefore is not expected to bind to Prb. Gankyrin was incorporated as a fusion to CspGFP to replace Pdar. Assaying binding between Prb and Gankyrin using split-spGFP generated no appreciable change in GFP fluorescence for up to 24 hours by flow cytometry **Figure 2.7B**. Thus, split-spGFP appears to dramatically increase the observable signal for protein-protein binding without compromising the negligible background of the original assay.

2.7 Conclusions

The split-GFP assay is a technically straight forward method to evaluate protein-protein

binding interactions *in vivo*, however, it's widespread application is stifled by poor signal to background ratio and limited utility at biologically relevant temperatures[16]. Previous work by Magliery and coworkers looked to remedy this by applying a variant of GFP with better folding kinetics to this system, creating split-frGFP[17]. In the approach presented here, we successfully leveraged the superior solubility and stability properties of supercharged proteins[13] to further improve the performance and broaden the utility of the assay. The resulting split-spGFP assay provided more efficient reassembly then previous variants under all conditions tested here. In future applications, split-spGFP fragments may provide a distinct advantage in evaluating PPIs between poorly soluble or inherently unstable proteins.

Split-spGFP performs very well at physiological temperature, enabling the study of interactions in a biologically relevant context. This is an important parameter for identifying ligands that effect a desirable change in a biological system. These results also demonstrate that split-spGFP can be applied to measuring PPIs of globular proteins. Collectively, this work further expands the potential binding systems that are amenable to analysis by split-GFP.

2.8 Methods

Preparation and expression of split-GFP fragments The plasmids N-sg100GFP-Z-pET11a and Z′-C-sg100GFP-pMRBAD were kindly provided by Lynn Regan's lab. N-sg100GFP-Z-pET11a encodes the N-fragment of split-sg100GFP fused to a leucine zipper, Z. Expression of the chimera is under the control of the T7 promoter and is induced by addition of

IPTG. pET11a holds an ampicillin resistance gene for selection in *E.coli*. Z′-C-sg100GFP-pMRBAD encodes the complementary leucine zipper, Z′, fused upstream of the C-fragment of split-sg100GFP. Expression is controlled by the Arabinose promoter and is induced by addition of (L)-arabinose. pMRBAD holds a kanamycin resistance gene for orthogonal selection from the pET11a plasmid in *E.coli*.

To obtain *E.coli* harboring both plasmids, approximately 20 ng of N-sg100GFP-Z-pET11a was first mixed with one aliquot of One Shot BL21 Star (DE3) chemically competent *E.coli* (Invitrogen). Heat shock transformation was carried out following the manufacturers instructions followed by plating of the rescued cells on LB ampicillin agar plates. The next morning, surviving colonies were picked and cultured in 5 mL of LB ampicillin. These cells were then made electrocompetent via standard protocols[22]. The complementary plasmid, Z′-C-sg100GFP-pMRBAD, was then electroporated into the competent *E.coli* cells already containing N-sg100GFP-Z-pET11a using the standard methods on a BioRad GenePulser Xcell electroporator. The rescued cells were plated on LB agar plates containing both ampicillin and kanamycin to concomitantly select for both plasmids. The next morning, surviving colonies were picked into LB supplemented with ampicillin and kanamycin.

An overnight culture of BL21 (DE3) *E.coli* harboring both N-sg100GFP-Z-pET11a and Z′-C-sg100GFP-pMRBAD plasmids was inoculated to an OD_{600} of < 0.1 in 10 mL. The culture was allowed to grow at 37 °C with shaking (200 RPM) until reaching an OD_{600} of ~0.6. At this time expression of both chimeras was induced by adding IPTG to a working concentration of

100 μM and (L)-arabinose to a working concentration of 0.2%. The culture was then incubated under the appropriate conditions as indicated.

Flow cytometry analysis of split-GFP reassembly After the proper induction and incubation, cells were prepared for analysis by centrifugation at 5000 RPM for 5 minutes, followed by resuspension in phosphate buffered saline (PBS). The GFP chromophore was excited by a solit-state iCyt 488 nm laser on a MoFlo flow cytometer (Dako Colorado, Inc.) to detect fluorescence. Flow cytometry data was analyzed using FloJo software (Tree Star data analysis software). Similar procedures were followed for the other constructs listed in this work.

Cloning of constructs used in this work The amino acid sequence of +36 GFP was kindly provided by David Liu's lab. Residues 1-157 of superpositive GFP were assembled using a common overlap PCR technique to construct N-spGFP. Similarly, residues 158-238 were cloned to make C-spGFP. N-sg100GFP was excised from pET11a by digestion with NdeI and XhoI followed by agarose gel purification. Following digestion with NdeI and XhoI, the N-spGFP PCR amplicon was ligated (*Quick Ligation Kit*, New England Biolabs) into pET11a upstream of leucine zipper Z, to create N-spGFP-Z-pET11a. A similar workflow was applied to assemble the construct Z′-C-spGFP-pMRBAD.

Plasmids encoding N-frGFP-Z-pET11a and Z′-C-frGFP-pMRBAD were kindly provided by Thomas Magliery's lab. Plasmids encoding Prb and Pdar were kindly provided by David Baker's lab. The constructs N-spGFP-Prb-pET11a, Pdar-C-spGFP-pMRBAD, and the equivalent split-frGFP plasmids were assembled using common molecular cloning techniques. Briefly, a Prb insert was prepared by PCR amplification using a 5′-primer containing an XhoI site

(5′-CCGCTCGAGGGCAGCACCCGTCCG-3′) and a 3′-primer containing a BamHI site (5′-ATATGGATCCTTACTATTTTTCGCCCA-GCAGGC-3′), followed by digestion with the appropriate restriction enzymes (NEB). The receiving vector was prepared by digesting N-spGFP-Z-pET11a with NdeI and XhoI restriction enzymes to remove leucine zipper Z. The two agarose gel purified components were ligated and transformed into One Shot TOP10 chemically competent *E.coli* (Invitrogen). DNA sequencing of the construct confirmed successful molecular cloning. A similar workflow was applied to assemble Pdar-C-spGFP-pMRBAD, N-frGFP-Prb-pET11a , and Pdar-C-frGFP-pMRBAD.

2.9 Proteins Used In This Work

NspGFP-linker-Z (shown with the leucine zipper)
nucleic acid sequence
ATGGGTCATCACCACCACCATCACGGTGGCGCTAGCAAAGGTGAACGTCTGTTT
CGTGGTAAAGTACCGATCTTAGTGGAATTAAAGGGCGACGTGAACGGTCATAAA
TTTAGCGTGCGCGGCGAAGGCAAAGGTGACGCTACCCGTGGTAAATTGACCCTG
AAGTTTATTTGCACAACAGGCAAATTACCCGTTCCGTGGCCCACCTTAGTGACCA
CCCTGACCTATGGCGTTCAGTGCTTCAGTCGTTACCCTAAACATATGAAACGTC
ACGATTTTTTCAAATCAGCCATGCCTAAAGGATATGTTCAAGAGCGTACAATCAG
CTTCAAGAAGGATGGCAAATATAAAACGCGTGCGGAAGTGAAATTTGAAGGCCGC
ACACTAGTAAATCGTATCAAACTGAAAGGTCGTGACTTCAAAGAAAAAGGCAACA
TTTTAGGCCATAAACTGCGTTATAACTTTAATTCTCATAAGGTGTATATTACGGC
CGATAAACGCGGTGGCTCTGGCTCTGGCTCGAGCGCCCTCAAAAAAGAATTGCA
GGCAAACAAAAAAGAACTTGCGCAGCTGAAGTGGGAGTTACAAGCTCTGAAAAAG
GAACTGGCGCAGTAA

amino acid sequence
MGHHHHHHGGASKGERLFRGKVPILVELKGDVNGHKFSVRGEGKGDATRGKLTLKF
ICTTGKLPVPWPTLVTTLTYGVQCFSRYPKHMKRHDFFKSAMPKGYVQERTISFKKDG

KYKTRAEVKFEGRTLVNRIKLKGRDFKEKGNILGHKLRYNFNSHKVYITADKRGGSGS
GSSALKKELQANKKELAQLKWELQALKKELAQ

Z′-linker-CspGFP (shown with the leucine zipper)
nucleic acid sequence:
ATGGCAAGCGAGCAGCTGGAAAAGAAGTTACAAGCCCTGGAGAAAAAACTTGCTC
AGCTGGAATGGAAAAACCAAGCATTGGAAAAAAAACTCGCGCAGACGTCGGGTG
GAAGCGGTAAGAATGGTATCAAGGCAAAATTCAAAATTCGCCATAACGTGAAAGA
CGGCAGCGTTCAATTAGCGGATCATTATCAACAAAACACGCCGATTGGTCGCGG
GCCTGTACTGTTACCTCGCAACCACTACCTGAGCACCCGTTCTAAACTGAGCAA
AGATCCGAAAGAAAAACGCGATCACATGGTTCTGTTAGAATTCGTGACCGCTGCA
GGCATTAAGCACGGACGCGACGAACGCTACAAGTAATGTACAACTAA

amino acid sequence
MASEQLEKKLQALEKKLAQLEWKNQALEKKLAQTSGGSGKNGIKAKFKIRHNVKDGS
VQLADHYQQNTPIGRGPVLLPRNHYLSTRSKLSKDPKEKRDHMVLLEFVTAAGIKHGR
DERYK

Nsg100GFP-linker
nucleic acid sequence
ATATGGCTAGCCATCACCACCATCACCATGGCGCGAGCAAAGGAGAAGAACTCT
TCACTGGAGTTGTCCCAATTCTTGTTGAATTAGATGGTGATGTTAACGGCCACA
AGTTCTCTGTCAGTGGAGAGGGTGAAGGTGATGCAACATACGGAAAACTTACCC
TGAAGTTCATCTGCACTACTGGCAAACTGCCTGTTCCATGGCCAACACTAGTCA
CTACTCTGTGCTATGGTGTTCAATGCTTTTCAAGATACCCGGATCATATGAAACG
GCATGACTTTTTCAAGAGTGCCATGCCCGAAGGTTATGTACAGGAAAGGACCAT
CTTCTTCAAAGATGACGGCAACTACAAGACACGTGCTGAAGTCAAGTTTGAAGGT
GATACCCTTGTTAATAGAATCGAGTTAAAAGGTATTGACTTCAAGGAAGATGGCA
ACATTCTGGGACACAAATTGGAATACAACTATAACTCACACAACGTTCCCATCAT
GGCAGACAAACAAGGTGGCTCTGGCTCTGGCTCGAGCAATCCCGGGAATTAAG

amino acid sequence
MASHHHHHHGASKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGKLTLKFI
CTTGKLPVPWPTLVTTLCYGVQCFSRYPDHMKRHDFFKSAMPEGYVQERTIFFKDDG
NYKTRAEVKFEGDTLVNRIELKGIDFKEDGNILGHKLEYNYNSHNVPIMADKQGGSGS
GSSNPGN
linker-Csg100GFP

nucleic acid sequence
CCATGGCTAATGCATGCAATGGGACGTCGGGTGGAAGCGGTAAGAATGGAATCA
AAGTGAACTTCAAGACCCGCCACAACATTGAAGATGGAAGCGTTCAACTAGCAGA
CCATTATCAACAAAATACTCCAATTGGCGATGGCCCTGTCCTTTTACCAGACAAC
CATTACCTGTCCACACAATCTGCCCTTTCGAAAGATCCCAACGAAAAGAGAGACC
ACATGGTCCTTCTTGAGTTTGTAACAGCTGCTGGGATTACACATGGCATGGATG
AACTGTACAACTAAG

amino acid sequence
MANACNGTSGGSGKNGIKVNFKTRHNIEDGSVQLADHYQQNTPIGDGPVLLPDNHY
LSTQSALSKDPNEKRDHMVLLEFVTAAGITHGMDELYN

NfrGFP-linker

nucleic acid sequence
TGGCGAGCCATCATCATCATCATCATGGCGTGAGCAAAGGCGAAGAACTGTTTA
CCGGCGTGGTGCCGATTCTGGTGGAACTGGATGGCGATGTGAACGGCCATAAAT
TTAGCGTGAGCGGCGAAGGCGAAGGCGATGCGACCTATGGCAAACTGACCCTGA
AATTTATTTGCACCACCGGCAAACTGCCGGTGCCGTGGCCGACCCTGGTGACCA
CCCTGACCTATGGCGTGCAGTGCTTTAGCCGCTATCCGGATCATATGAAACAGC
ATGATTTTTTTAAAAGCGCGATGCCGGAAGGCTATGTGCAGGAACGCACCATTA
GCTTTAAAGATGATGGCAACTATAAAACCCGCGCGGAAGTGAAATTTGAAGGCGA
TACCCTGGTGAACCGCATTGAACTGAAAGGCATTGATTTTAAAGAAGATGGCAAC
ATTCTGGGCCATAAACTGGAATATAACTATAACAGCCATAACGTGTATATTACCG
CGGATAAACAG

amino acid sequence
MASHHHHHHGVSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGKLTLKFI
CTTGKLPVPWPTLVTTLTYGVQCFSRYPDHMKQHDFFKSAMPEGYVQERTISFKDDG
NYKTRAEVKFEGDTLVNRIELKGIDFKEDGNILGHKLEYNYNSHNVYITADKQ

linker-CfrGFP

nucleic acid sequence
AAAAACGGCATTAAAGCGAACTTTAAAATTCGCCATAACATTGAAGATGGCAGCG
TGCAGCTGGCGGATCATTATCAGCAGAACACCCCGATTGGCGATGGCCCGGTGC
TGCTGCCGGATAACCATTATCTGAGCACCCAGAGCGCGCTGAGCAAAGATCCGA
ACGAAAAACGCGATCATATGGTGCTGCTGGAATTTGTGACCGCGGCGGGCATTA
CCCATGGCATGGATGAACTGTATAAA

amino acid sequence
KNGIKANFKIRHNIEDGSVQLADHYQQNTPIGDGPVLLPDNHYLSTQSALSKDPNEKR
DHMVLLEFVTAAGITHGMDELYK

Prb

nucleic acid sequence
GGCAGCACCCGTCCGATTGATGGTCTGACGGACGAAGATATCCGCGAGATTCTG
ACTCGTTATAAAAAGATCGCTCTGGTGGGGGCCTCTCCGAAACCGGAACGCGAC
GCAAATATTGTTATGAAATATCTGCTGGAGCATGGCTATGATGTCTACCCGGTAA
ACCCGAATTATGAAGAAGTGCTGGGCCGCAAGTGCTACCCGAGTGTTCTGGACA
TCCCGGATAAAATTGAGGTCGTAGATCTGTTTGTGAACCCGGCTAAAGCGTGGC
GCTTCGTTGTCTATGCCATCAAGAAAGGGGCAAAAGTGGTATGGTTTCAGTACA
ACACATATTATCCGCTGGCTGCGCGTCAAGCCAAGGAAGCAGGCCTGATTATCG
TTGCTAATCGCTGTATGATGCGTGAGCACGAACGCCTGCTGGGCGAAAAA

amino acid sequence
GSTRPIDGLTDEDIREILTRYKKIALVGASPKPERDANIVMKYLLEHGYDVYPVNPNYEE
VLGRKCYPSVLDIPDKIEVVDLFVNPAKAWRFVVYAIKKGAKVVWFQYNTYYPLAARQA
KEAGLIIVANRCMMREHERLLGEK

Pdar

nucleic acid sequence
AGCGATCTGGGTAAAAAGCTGCTGGAAGCAGCCGCGGCCGGCCAAGATGATGAG
GTGCGTATTCTGATGGCGAATGGGGCCGATGTTAACGCAACCGACGACGATGGC
CTGACTCCGCTGCACCTGGCGGCTGCAAACGGGCAACTGGAAATCGTAGAGGTA
CTGCTGAAAAATGGCGCCGATGTGAACGCTTCTGATAGTGCGGGTATTACTCCG
CTGCACCTGGCCGCTTATGACGGCCATCTGGAGATTGTCGAAGTCCTGCTGAAG
CACGGGGCTGACGTTAATGCGTACGACCGCGCCGGGTGGACACCGCTGCACCT
AGCAGCGCTGAGTGGCCAACTGGAGATTGTGGAAGTTCTGCTGAAACACGGCGC
AGATGTCAACGCCCAAGACGCACTGGGCCTGACCGCGTTTGATATCTCGATTAA
TCAAGGTCAGGAAGATCTGGCAGAGATCCTGCAA

amino acid sequence
SDLGKKLLEAAAAGQDDEVRILMANGADVNATDDDGLTPLHLAAANGQLE
IVEVLLKNGADVNASDSAGITPLHLAAYDGHLEIVEVLLKHGADVNAYDR
AGWTPLHLAALSGQLEIVEVLLKHGADVNAQDALGLTAFDISINQGQEDLAEILQ

Gankyrin

nucleic acid sequence

ATGGAGGGGTGTGTGTCTAACATAATGATCTGTAACCTGGCCTACAGTGGGAAG
CTGGATGAGTTGAAGGAGCGCATTTTGGCTGATAAATCTCTGGCTACTAGAACT
GATCAGGACAGCAGAACAGCTTTGCACTGGGCATGCTCAGCTGGCCATACAGAA
ATTGTTGAATTCTTGCTGCAACTTGGAGTGCCAGTGAATGATAAAGATGACGCA
GGTTGGTCTCCTCTTCATATTGCTGCCTCCGCTGGCCGGGATGAGATTGTAAAA
GCCCTTCTGGTGAAAGGTGCACATGTGAATGCTGTCAATCAAAACGGCTGCACT
CCACTCCATTATGCAGCTTCGAAGAATAGGCATGAGATTGCTGTTATGTTACTAG
AAGGTGGGGCTAACCCAGATGCGAAGGACCATTACGATGCTACAGCAATGCACC
GGGCAGCAGCCAAGGGTAACTTGAAGATGGTTCACATCCTTCTGTTCTACAAAG
CATCCACAAACATCCAAGACACTGAGGGTAACACTCCTCTACACTTAGCCTGTGA
TGAAGAGAGAGTGGAAGAGGCAAAATTTCTGGTGACTCAAGGAGCAAGTATTTAC
ATTGAGAATAAAGAAGAAAAGACACCCCTGCAAGTGGCCAAAGGGGGCCTGGGT
TTAATACTCAAGAGACTAGCAGAAGGTGAAGAGGCTTCTATGTAA

amino acid sequence

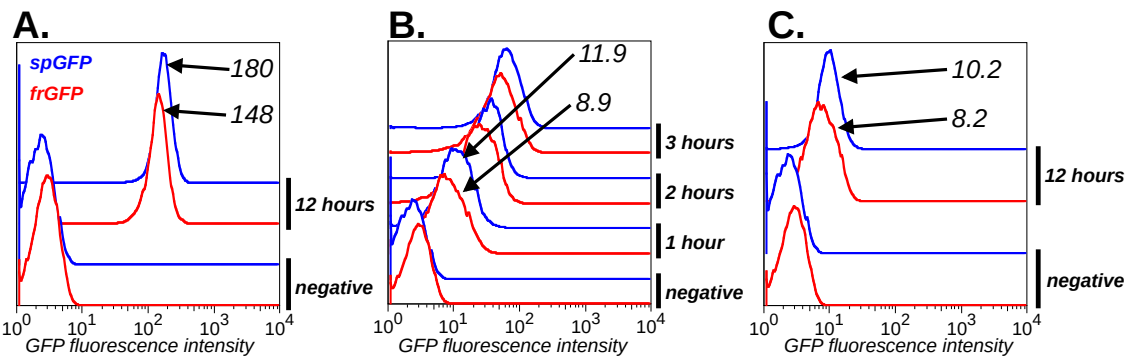
MEGCVSNIMICNLAYSGKLDELKERILADKSLATRTDQDSRTALHWACSAGHTEIVEFL
LQLGVPVNDKDDAGWSPLHIAASAGRDEIVKALLVKGAHVNAVNQNGCTPLHYAASK
NRHEIAVMLLEGGANPDAKDHYDATAMHRAAAKGNLKMVHILLFYKASTNIQDTEG
NTPLHLACDEERVEEAKFLVTQGASIYIENKEEKTPLQVAKGGLGLILKRLAEGEEASM

2.10 Supplemental Data

Analysis of flow cytometry data comparing split-frGFP and split-spGFP

All flow cytometry data was analyzed using FloJo software.

Figure S2.1. Mean cell fluorescence values are provided for **Figure 2.5**.

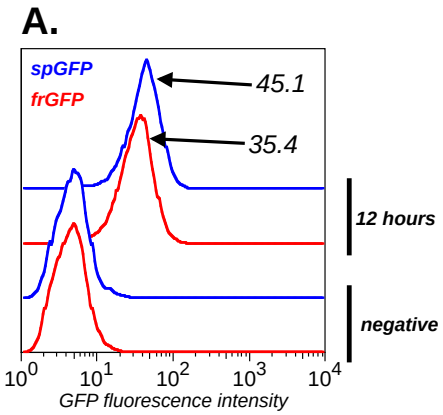


180 – 148 = 32
32/148 = 0.22
22% increase in spGFP cell fluorescence

11.9 – 8.9 = 2
2/8.9 = 0.34
34% increase in spGFP cell fluorescence

10.2 - 8.2 = 2
2/8.2 = 0.24
24% increase in spGFP cell fluorescence

Figure S2.2 Mean cell fluorescence values are provided for **Figure 2.8A**



45.1 – 35.4 = 9.7
9.7/35.4 = 0.27
27% increase in spGFP cell fluorescence

REFERENCES

(1) Yang, M.; Wu, Z.; Fields, S. *Nucleic Acids Res.* **1993**, *23*, 1152.

(2) Stynen, B.; Tournu, H.; Tavernier, J.; Van Dijck, P. *Microbiol. Mol. Biol. Rev.* **2012**, *76*, 331.

(3) Remy, I.; Michnick, S. W. *Biotechniques* **2005**, *42*, 137.

(4) Michnick, S. W.; Ear, P. H.; Landry, C.; Malleshaiah, M. K.; Messier, V. In *Antibody Engineering*; Humana Press: Totowa, NJ, 2011; Vol. 756, pp. 395–425.

(5) Gietz, R. D.; Woods, R. A. *BioTechniques* **2001**, *30*, 816.

(6) Sambrook, J.; Fritsch, E.; Maniatis, T. *Molecular cloning*; Cold Spring Harbor Labratory Press, 1989.

(7) Galarneau, A.; Primeau, M.; Trudeau, L.-E.; Michnick, S. W. *Nat. Biotechnol.* **2002**, *20*, 619.

(8) Broome, A.; Bhavsar, N.; Ramamurthy, G.; Newton, G.; Basilion, J. P. *Mol Pharma* **2010**, *93*, 1375.

(9) Remy, I.; Michnick, S. W. *Proc. Natl. Acad. Sci. USA* **1998**, *96*, 5394.

(10) Van Zeebroeck, G.; Kimpe, M.; Vandormael, P.; Thevelein, J. M. *PLoS One* **2011**, *6*, e24275.

(11) Wilson, C. G. M.; Magliery, T. J.; Regan, L. *Nat. Methods* **2004**, *1*, 255.

(12) Magliery, T. J.; Wilson, C. G. M.; Pan, W.; Mishler, D.; Ghosh, I.; Hamilton, A. D.; Regan, L. *J. Am. Chem. Soc.* **2005**, *127*, 146.

(13) Lawrence, M. S.; Phillips, K. J.; Liu, D. R. *J. Am. Chem. Soc.* **2007**, *129*, 10110.

(14) Thompson, D. B.; Cronican, J. J.; Liu, D. R. *Engineering and identifying supercharged proteins for macromolecule delivery into mammalian cells.*; 1st ed.; Elsevier Inc., 2012; Vol. 503, pp. 293–319.

(15) McNaughton, B. R.; Cronican, J. J.; Thompson, D. B.; Liu, D. R. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 6111.

(16) Magliery, T. J.; Regan, L. *Methods Biochem. Anal.* **2005**, *47*, 391.

(17) Sarkar, M.; Magliery, T. J. *Mol. Biosyst.* **2008**, *4*, 599.

(18) Jackrel, M. E.; Cortajarena, A. L.; Liu, T. Y.; Regan, L. *ACS Chem. Biol.* **2010**, *5*, 553.

(19) Karanicolas, J.; Corn, J.; Chen, I.; Joachimiak, L. *Mol. Cell* **2011**, *42*, 250.

(20) Ghosh, I.; Hamilton, A. D.; Regan, L.; *J Am Chem Soc* **2000**, *122*, 5658.

(21) Nakamura, Y.; Nakano, K.; Umehara, T.; Kimura, M.; Hayashizaki, Y.; Tanaka, A.; Horikoshi, M.; Padmanabhan, B.; Yokoyama, S. *Structure* **2007**, *15*, 179.

(22) Sidhu, S.; Feld, B.; Weiss, G. *Protein Eng. Protoc.* **2007**, *353*, 205.

CHAPTER THREE
TOWARD THE DEVELOPMENT OF SYNTHETIC RNA RECOGNITION MOTIFS: EXAMINING THE LIMITS AND DICTATES OF PROTEIN-RNA INTERACTIONS INVOLVING U1A AND U1HPII MUTANTS

3.1 Introduction

A multitude of sequence diverse RNA hairpins have been implicated in human disease[1–3]. Thus, interrogating and manipulating disease-relevant interactions and biochemistry of RNA molecules has important implications for studying and correcting RNA-dependent diseases. Fundamentally, modulation of disease-relevant targets, including RNA, is a molecular recognition problem. Researchers must first identify reagents capable of potently and selectively recognizing the target – often in a complex solution. Thus, the development of probes for potent and selective recognition of a target RNA molecule is critical. Small molecule-based reagents for this purpose have remained elusive. Intrinsically, "small" molecules (< 500 Da) tend to lack the chemical and structural complexity necessary to identify the chemically uniform and highly charged surfaces of nucleic acid polymers in a selective manner[4]. This has prompted the exploration of larger molecular scaffolds capable of enveloping a broader target surface area and, consequently, providing more selective interaction with a single nucleic acid molecule.

The development of polyamide molecules that potently and selectively recognize a specific sequence of double-stranded B-form DNA (dsDNA) **Figure 1.9** is regarded as a seminal achievement in modern chemical biology and molecular recognition[5–7]. Polyamide's

programmable selectivity for virtually any dsDNA sequence is largely attributed to the (generally) uniform three-dimensional structure of dsDNA – which typically adopts a B-form helix. Development of an analogous molecular scaffold for programmed recognition of virtually any RNA sequence is unlikely, given the structurally diverse architectures that RNA can adopt. Thus, while sequence-defined recognition can be solved – from basic principles – the design of synthetic reagents capable of analogous recognition of folded RNAs – again, from basic principles – remains an unsolved and particularly challenging problem.

A potentially better solution might be found in the development of medium- to high-molecular weight molecules. In contrast to their small molecule cousins, these reagents have the surface area – and potentially the structure – to potently and selectively recognize an RNA with a particular sequence and shape. Notable progress toward this goal includes Nielsen's peptide nucleic acids (PNAs) (~1500-2000 Da) for binding single-stranded RNA[8–10] and Disney's poly-aminoglycosides for binding repeat stem loops[11–13] **Figure 1.12.** However, these molecules have their own drawbacks. Primary among these is the fact that these relatively large molecules must be prepared using chemical synthesis, which is relatively laborious and expensive. Also, PNAs still operate within the confines of sequence-defined recognition derived from already established basic principles. Additionally, Disney's pseudo-polymeric molecules utilize molecular modules that are known to exhibit general affinity for DNA and RNA (such as aminoglycosides or molecules that resemble Hoescht stain). Thus, questions remain about how selective these molecules are relative to DNA or for a particular RNA sequence. Therefore,

despite these advances, the development of reagents that appreciate folded RNA structures in three-dimensional space during recognition persists as a technically challenging goal.

An alternative approach is to borrow an RNA-binding protein scaffold from Nature, and alter (as opposed to merely broaden) its RNA binding specificity. This approach is promising for a number of reasons. First, in contrast to relatively well-studied RNA-binding small molecules, such as aminoglycosides or aromatic stains such as Hoescht, which bind RNA with relatively low affinity (K_D > 50-100 μM)[14,15], many RNA-binding proteins recognize their RNA target with exceptional affinity, which can be as good as $K_D \sim 10^{-11}$ M[16,17]. Second, protein scaffolds are easily modified, to diversify their chemical functionality, and combinatorially screened for improved function (affinity), such as recognition of a specific RNA sequence **Figure 1.17**. This is in contrast to small molecule discovery, which requires laborious and time consuming chemical modification and individual screening for the desired activity. The remaining question is: which RNA-binding protein scaffold should we use to develop new proteins capable of binding folded RNA molecules?

3.2 The RNA Hairpin-Binding Properties of the U1A RRM

The N-terminal region of the U1 small nuclear ribonucleoprotein (U1A) is perhaps the best studied RNA Recognition Motif (RRM) protein[18,19] **Figure 3.1A**. U1A fulfills many of the

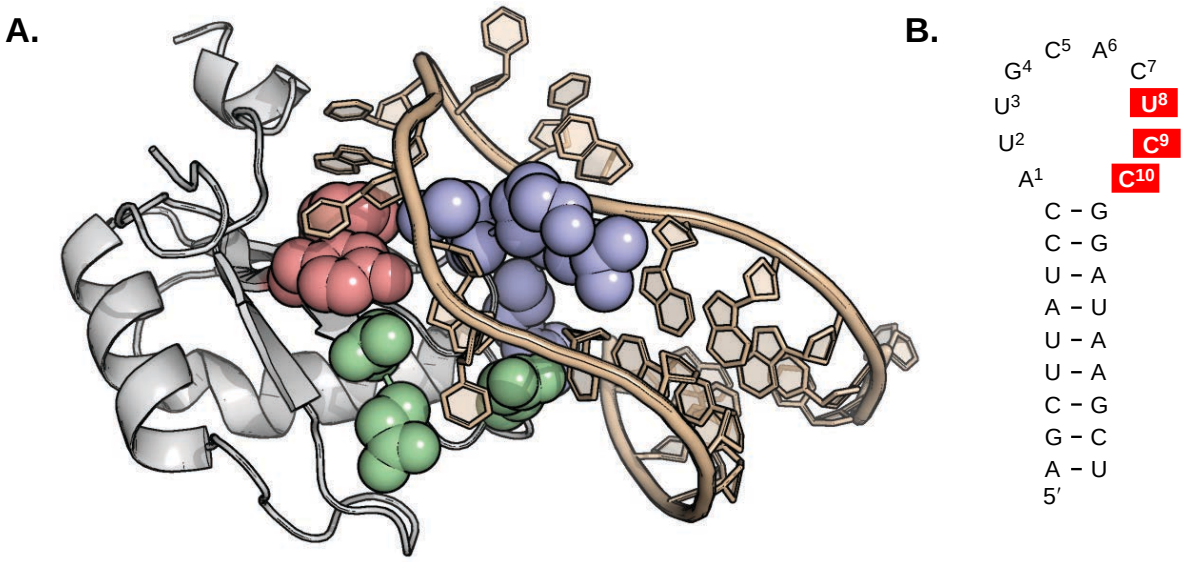


Figure 3.1 A The protein-RNA binding interaction of U1A (*silver*) and U1hpII RNA (*wheat*). Putative RNA binding modules relevant to this study are highlighted as spheres on U1A: The β2-β3 loop (*blue*), the β1-α1 loop (*green*), and aromatic residues on the β-sheet face (*red*). **B** Nucleotide sequence and Watson-Crick binding interactions in the hairpin structure of U1hpII RNA. Hairpin loop residues are numbers as they are referenced in the text and "spacer" nucleotides are highlighted in red. (PDB code: 1URN)

requirements to serve as a scaffold for developing new RRMs that bind disease-relevant folded RNAs. First, U1A's cognate RNA (U1 hairpin II RNA, U1hpII) has a defined folded secondary structure that forms a hairpin **Figure 3.1B**. This suggests that the U1A protein scaffold is predisposed to recognize this general structural class of hairpin RNAs. Second, U1A selectively binds U1hpII with incredibly tight affinity (K_D, ~30 pM)[20], suggesting that this scaffold is potentially a very good solution to the specific challenge of RNA hairpin recognition. Lastly, as described below, we know a lot about the molecular requirements for the U1A/U1hpII complex – detailed structural and biophysical characterization of this complex has been reported. Thus, our knowledge of the putative RNA-binding "modules" or sub-domains within U1A **Figure 3.1A**, *highlighted as spheres*, allows us to thoughtfully modify them. Rational design and / or molecular evolution applied to these sub-domains could potentially result in new U1A-derived

proteins that potently and selectively bind a new disease-relevant RNAs that have a similar hairpin tertiary structure **Figure 1.18**.

Binding between U1A and U1hpII proceeds through a "lure and lock" mechanism[20], wherein positively charged surface residues on U1A "lure" the RNA into close proximity – followed by non-selective and sequence-selective interactions between the RNA and U1A sub-domains. Foremost among these sub-domains, with respect to stabilizing the interaction, is the β-sheet face of U1A, **Figure 3.1A**, *red*. Upon engaging the RNA, aromatic residues Tyr13 and Phe56 π-stack with nucleobases C5 and A6, respectively[18,21–23] **Figure 3.2A**. Laird-Offringa and coworkers have demonstrated that mutations to these residues cause a dramatic decrease in U1hpII binding affinity[22]. While these interactions do not play a strong role in sequence *selective* recognition of U1hpII, they clearly contribute significantly to the stability of the complex.

Previous researchers have also shown that residues on a second U1A sub-domain, the β1-α1 loop **Figure 3.1A** *green* also play a significant role in U1hpII [21,2425]. Work from the Baranger lab has provided evidence that Asn15, Asn16, and Glu19 contact U1hpII through a sequence-dependent hydrogen bond network with bases U2, U3, and G4[26] **Figure 3.2B**. Collectively, each of these productive contacts "lock" the RNA into place, resulting in the high binding affinity observed in this reaction.

A third U1A sub-domain, the β2-β3 loop **Figure 3.1A** *blue*, is a prominent structural feature on the RNA binding face on the RRM[27,2829]. The β2-β3 loop protrudes through the

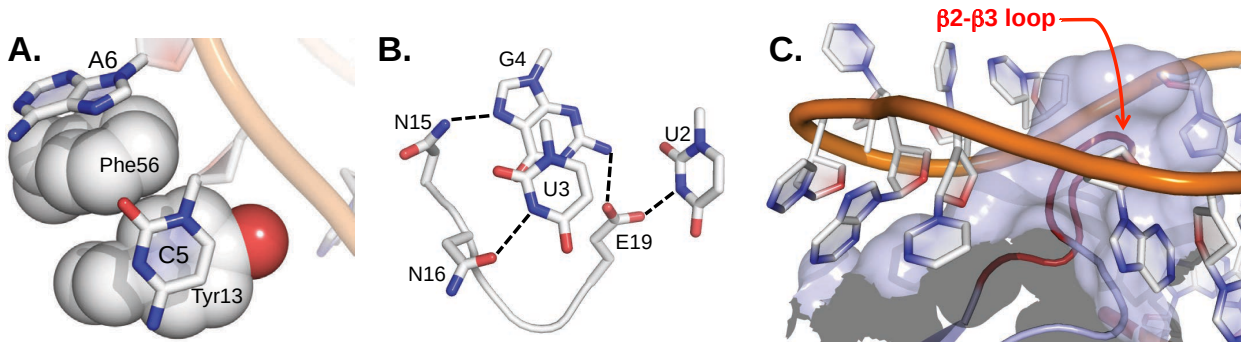


Figure 3.2 Amino acid-nucleotide interactions that are important for stabilizing the U1A-U1hpII interface. **A** π-stacking interactions of aromatic residues on the β-sheet face of U1A, Tyr13 and Phe56, with bases Cytosine 5 and Adenosine 6, respectively. **B** Hydrogen bonding interactions formed between residues on the α1-β1 loop of U1A with bases of U1hpII that impart sequence specificity to the binding interaction. **C** The β2-β3 loop of U1A (*red*) protrudes up through the hairpin loop of U1hpII, likely helping to immobilize the RNA on the protein scaffold. (PDB code: 1URN)

hairpin loop of U1hpII, like a finger into a glove **Figure 3.2C**. This shape recognition by the β2-β3 loop fits U1hpII loop bases in proximity to putative RNA-binding modules, or sub-domains, described above. Notably, none of the side chain residues in the β2-β3 loop itself directly interact with hairpin loop bases in U1hpII **Figure 3.3A**. This region likely imposes steric restrictions on the bound RNA, further stabilizing productive contacts.

One potential pit-fall of U1A as a general RRM scaffold for disease-relevant RNA hairpins is the large size of the loop region on its cognate RNA. In contrast to many disease-relevant RNA hairpins, which often have a loop of ~4-8 nucleotides (nt) in length, U1hpII has a 10-nt loop[18,25]. Such large loops are often the recognition site for RNA splicing proteins, and thus, are relatively under populated in natural RNA hairpin space. However, features of the U1A/U1hpII interactions suggested that proper modification to the U1A scaffold could generate a derivative RRM that binds hairpins with smaller loops. First, none of the residues in the β2-β3 loop make productive base contacts when bound to U1hpII **Figure 3.3A**, but rather

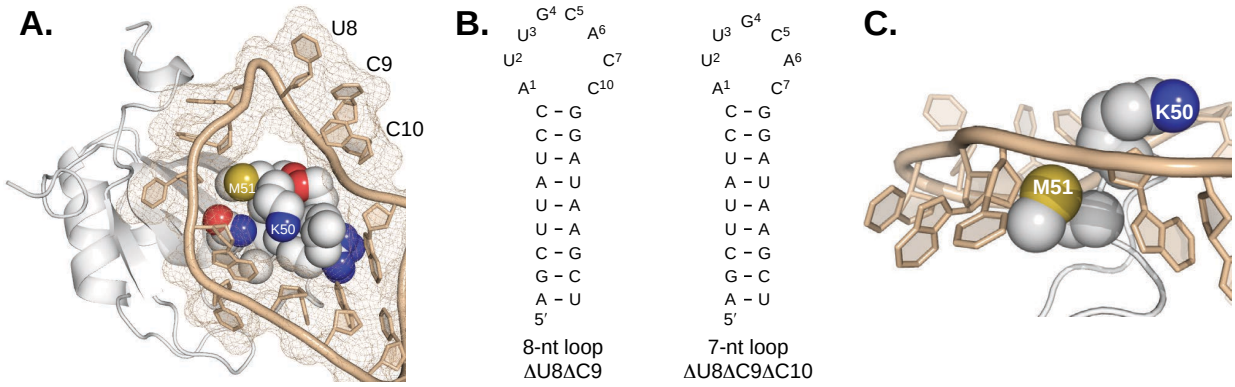


Figure 3.3 A Arial perspective of **Figure 1a** with β2-β3 loop residues highlighted as spheres. "Spacer" nucleotides U8, C9, and C10 do not interact with U1A in the native binding interaction. **B** Depiction of the primary and secondary structure of the truncated U1hpII variants with loop nucleotides numbered. **C** Close up perspective of the β2-β3 loop showing that neither Lys50 nor Met51 interact with U1hpII in the native binding interaction. (PBD code: 1URN)

sterically immobilize the RNA on the RRM scaffold[28]. It therefore may be possible to reduce the bulk of this steric peg, making space for smaller hairpin loops. Secondly, U1hpII bases U8, C9, and C10 **Figure 3.1B** *boxed in red* appear to be disordered in the crystal structure of the bound complex[18], suggesting they do not directly contribute to the U1A-U1hpII interface **Figure 3.3A**. As an extension of this empirical observation, Laird-Offringa and coworkers demonstrated the flexibility of this region in the RNA[30]. This suggests that the nucleotides indeed contribute minimal, if any, productive contacts with U1A and merely act as a spacer in the U1hpII hairpin loop. If these "spacer" nucleotides are removed entirely, can U1A still accommodate the smaller RNA hairpins, such as those found in most naturally occurring RNA hairpins?

To further investigate the variability of U1hpII hairpin loop size in the interaction with U1A, Laird-Offringa and coworkers measured the binding affinity of U1A with a truncated variant of U1hpII[28]. Removing bases U8 and C9 from U1hpII creates an 8-nt loop variant,

U1hpII ΔU8ΔC9 **Figure 3.3B** *left*. This deletion resulted in a 3000-fold decrease in complex affinity. However, a compensatory mutation to the β2-β3 loop of U1A afforded a double deletion mutant ΔK50ΔM51, that binds U1hpII and U1hpII ΔU8ΔC9 with similar affinities. Removal of Lys50 and Met51 **Figure 3.3C** reduces the steric bulk in the β2-β3 loop, which perhaps clears space to accommodate the smaller hairpin loop. These findings suggest that the resulting U1A mutant, ΔK50ΔM51, may be better suited to bind smaller hairpin loops than wtU1A, perhaps as a result of reduced steric bulk in the β2-β3 loop.

This is potentially significant for our goal of developing a general platform for RNA hairpin recognition for two primary reasons. First, ΔK50ΔM51 should retain the privileged RNA hairpin recognition properties of its parent scaffold, U1A. Second, ΔK50ΔM51 provides the added benefit of also accommodating smaller hairpin loop sizes, more commonly found in naturally occurring RNA hairpins. For these reasons, ΔK50ΔM51 therefore represents a compelling starting point to begin looking for new RRMs that target physiologically relevant hairpin RNAs involved in interesting or unique biochemistry.

3.3 Mapping the Interface of ΔK50ΔM51 with ΔU8ΔC9 U1hpII RNA

As previously stated, part of our interest in ΔK50ΔM51 is predicated on its inheritance of

the exceptional RNA hairpin recognition properties from the parent scaffold, U1A. For this reason, it is imperative to ensure that the binding mode of ΔK50ΔM51 in complex with smaller hairpin loop RNAs is in agreement the native binding interaction. Additionally, future rational design or molecular evolution efforts applied to ΔK50ΔM51 will likely require a more complete understanding of this interface. To begin, we measured the affinity between ΔK50ΔM51 and the 8-nt loop variant of U1hpII, U1hpII ΔU8ΔC9, using fluorescence polarization **Figure 3.3B** *left*. This interaction gave a dissociation constant (K_D) of 4.0 (±0.4) x 10^{-6} **Table 3.1 entry 1**, which is consistent with the previously measured[28] K_D of 7 (±4) x 10^{-7}. This was encouraging as it validated a method to quantify this protein-RNA interaction, and other interactions involving ΔK50ΔM51 derivatives, with good throughput.

To correlate the binding mode of ΔK50ΔM51-U1hpII ΔU8ΔC9 to the native interaction we created a small but focused library of point mutations on the ΔK50ΔM51 scaffold. Mutated residues map to positions that are critical for binding in the native U1A-U1hpII complex **Figure 3.4A**. Aromatic residues on the β-sheet face of U1A are conserved among most RRM proteins and are generally important for RNA binding[31,32]. We evaluated the importance of these residues, Tyr13 and Phe56 **Figure 3.4A** *red*, in this new binding interface between a synthetic RRM, ΔK50ΔM51, and the 8-nt loop U1hpII ΔU8ΔC9, by disrupting their potential binding contribution. In the native U1A-U1hpII complex, Tyr13Gln and Phe56Ala mutations each dramatically disrupt binding affinity[22]. Introducing the analogous mutants into ΔK50ΔM51 individually decreased binding affinity for the 8-nt loop variant of U1hpII by ~134- and ~104-

Table 3.1 Binding Affinities for U1A ΔK50ΔM51 Mutants and Eight-Nucleotide Loop Variants of U1hpII RNA

entry	protein mutant	RNA mutant	K_D[a] (M)	ΔG° (kcal/mol)[c]	ΔΔG° (kcal/mol)[d]
1	U1A ΔK50ΔM51	8nt loop U1hpII	4.0(±0.5) x 10^{-6}	-7.35	-
2	Y13Q[b]	8nt loop U1hpII	536.9 (±106.5) x 10^{-6}	-4.45	2.9
3	F56A	8nt loop U1hpII	414.9(±81.0) x 10^{-6}	-4.61	2.74
4	N15A	8nt loop U1hpII	33.9(±2.0) x 10^{-6}	-6.09	1.26
5	N16A	8nt loop U1hpII	15.7(±1.0) x 10^{-6}	-6.54	0.81
6	E19A	8nt loop U1hpII	4.7(±0.5) x 10^{-6}	-7.26	0.09
7	U1A ΔK50ΔM51	8nt loop U1hpII G4A	11.8(±2.5) x 10^{-6}	-6.71	0.64

[a]The error for dissociation constant (K_D) is the standard deviation of three separate experiments. [b] All point mutants are derived from U1A ΔK50ΔM51 [c]ΔG is the free energy of the protein in complex with U1hpII-derived RNA calculated with the equation $\Delta G = -RT \ln K_D$. [d]ΔΔG is the difference in binding free energy between the complexes in entry 1 and the indicated mutants thereof.

fold, respectively **Table 3.1 entries 2** and **3**. This suggests that these residues play a similarly important role in RNA binding, as they do in U1A and most RRMs.

We next looked to interrogate the β1-α1 loop subdomain **Figure 3.4A** *green* of the ΔK50ΔM51 scaffold. As mentioned previously, residues in this region engage in a hydrogen bond network with U1hpII bases in the native interaction[26] **Figure 3.3B**. Asn15Ala and Asn16Ala mutations to ΔK50ΔM51 revealed ~8.5- and ~3.9-fold decreases in affinity for the 8-nt loop variant, respectively **Table 3.1 entries 4** and **5**. This is again consistent with previously reported SAR in U1A, in which Asn15Ala and Asn16Ala mutations reduced affinity by ~10- and ~6-fold for U1hpII, respectively. Collectively, this shows that a number of productive binding contacts are preserved in the non-natural interaction, suggesting a shared binding mode with the native complex.

Glu19, also on the β1-α1 loop, contributes significantly to the native binding interaction, evidenced by the ~50-fold decrease in affinity induced by a Glu19Ala mutation[21]. Interestingly, the identical mutation to ΔK50ΔM51 did not appreciably alter the interaction with the 8-nt loop

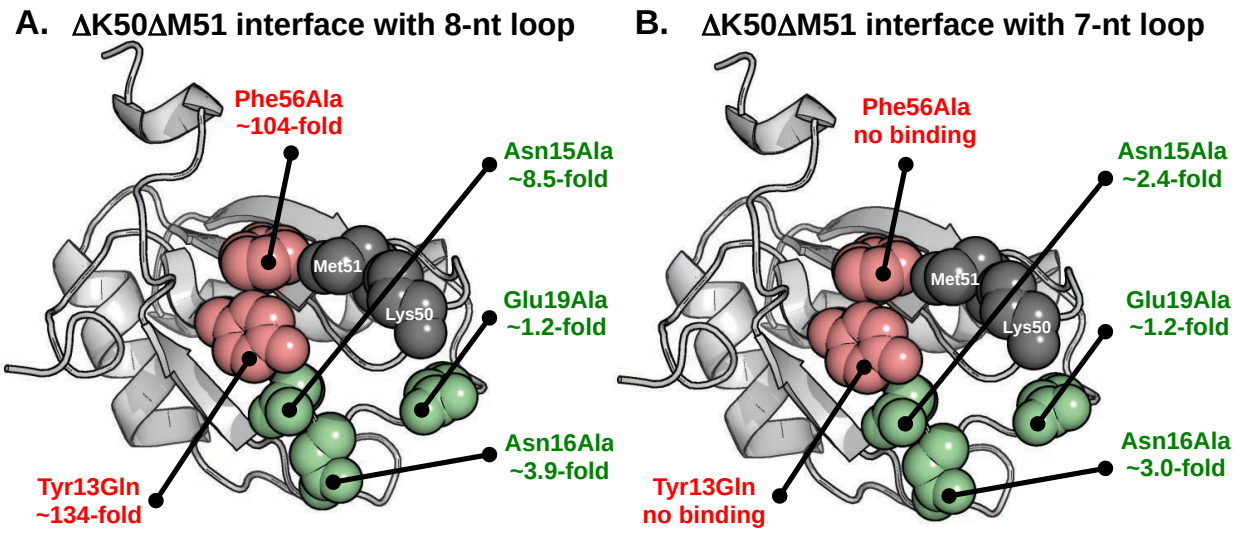


Figure 3.4 Relative decreases in affinity of ΔK50ΔM51 mutants in complex with either the **A** 8-nt loop variant or **B** 7-nt loop variant of U1hpII. Fold changes are relative to ΔK50ΔM51. The ΔK50ΔM51 residues examined here map onto putative RNA binding regions of U1A. Black spheres, Lys50 and Met51, represent the residues removed from U1A to make ΔK50ΔM51. (PDB code: 1URN)

variant **Table 3.1 entry 7**. In the native interaction, this residue and Asn15 each engage in a hydrogen bond with G4 in the U1hpII loop **Figure 3.2B**. Baranger and coworkers demonstrated the importance of this interaction by introducing an adenine base at this position, which led to a ~10^5 decrease in affinity[21]. To further probe the role of Glu19 in ΔK50ΔM51, we introduced this same G4A substitution to create U1hpII ΔU8ΔC9 G4U RNA. Binding affinity between ΔK50ΔM51 and the G4A substituted RNA decreased by only ~2.9 fold **Table 3.1 entry 8**, further suggesting a muted role for Glu19 in recognizing 8-nt hairpin loops.

3.4 Mapping the Interface of ΔK50ΔM51 with ΔU8ΔC9ΔC10 U1hpII RNA

Encouraged by similarities between the measured interaction and the native complex, we next looked to further explore ΔK50ΔM51's proclivity for binding smaller hairpin loop RNAs. As mentioned previously, three loop nucleotides of U1hpII, including C10, do not engage the protein surface when the RNA is bound **Figure 3.3A**. Removal of all three spacer nucleotides creates a 7-nt loop variant, U1hpII ΔU8ΔC9ΔC10 RNA **Figure 3.3B** *right*. To our delight, we found that ΔK50ΔM51 retains good affinity for this smaller loop variant (14.5 ±2.9 μM) **Table 3.2 entry 1**. This is significant because it broadens the functional range of loop sizes that ΔK50ΔM51 can accommodate with good affinity, relative to recently reported RNA-binding small molecules. As with the 8-nt loop, we then measured similar SAR for the interaction between ΔK50ΔM51 and the 7-nt loop variant of U1hpII **Figure 3.4B**. Again, Tyr13Gln and Phe56Ala mutations to ΔK50ΔM51 greatly destabilized this complex **Table 3.2 entries 2** and **3**, suggesting that ΔK50ΔM51 binds RNA in a manner similar to most RRMs and, specifically to its parent scaffold. Additionally, Asn15Ala and Asn16Ala mutations led to ~2.4-fold and ~3-fold lowering in affinity, respectively **Table 3.2 entries 4** and **5**. This is also consistent with data for the 8-nt loop variant and for the native interaction. Also consistent with the 8-nt loop data, but in contrast to the native interaction, Glu19Ala had little effect on binding to the 7-nt loop **Table 3.2 entry 6**. Introducing a G4A substitution into the 7-nt loop also had a negligible effect on the binding interaction **Table 3.2 entry 7**. These data set suggests that most of the productive contacts that drive the native binding affinity are also important for

Table 3.2 Binding Affinities for U1A ΔK50ΔM51 Mutants and Seven-Nucleotide Loop Variants of U1hpII RNA

entry	protein mutant	RNA mutant	K_D[a] (M)	ΔG° (kcal/mol)[c]	ΔΔG° (kcal/mol)[d]
1	U1A ΔK50ΔM51	7nt loop U1hpII[a]	14.5(±2.9) x 10^{-6}	-6.59	-
2	Y13Q[b]	7nt loop U1hpII	no binding	---	---
3	F56A	7nt loop U1hpII	no binding	---	---
4	N15A	7nt loop U1hpII	34.8(±2.7) x 10^{-6}	-6.07	0.52
5	N16A	7nt loop U1hpII	43.9(±2.3) x 10^{-6}	-5.94	0.65
6	E19A	7nt loop U1hpII	17.7(±3.5) x 10^{-6}	-6.47	0.12
7	U1A ΔK50ΔM51	7nt loop U1hpII G4A	21.2(±3.1) x 10^{-6}	-6.36	0.23

[a]The error for dissociation constant (K_D) is the standard deviation of three separate experiments. [b] All point mutants are derived from U1A ΔK50ΔM51 [c]ΔG is the free energy of the protein in complex with U1hpII-derived RNA calculated with the equation $\Delta G = -RT \ln K_D$. [d]ΔΔG is the difference in binding free energy between the complexes in entry 1 and the indicated mutants thereof.

stabilizing the association of ΔK50ΔM51 with 7-nt loops. The one exception identified here is that Glu19 is indifferent, which agrees with the 8-nt loop data set.

3.5 Conclusions

Based on previous findings, we proposed that the ΔK50ΔM51 scaffold may function as a template for generating new recognition elements that bind physiologically relevant RNA hairpins. This is contingent on two features of the scaffold: 1) ΔK50ΔM51 retains the enhanced RNA hairpin-binding properties of its parent scaffold, U1A, and 2) by virtue of reduced steric bulk in the β2-β3 loop, ΔK50ΔM51 is more accommodating of RNA hairpins with smaller loops.

In this study, we have shown that most of the contacts that drive the native binding interaction also contribute to the association of ΔK50ΔM51 with 8- and 7-nt loop hairpins. These data appear to be indicative of a shared binding mode between the native complex and

the interactions measured here. The Glu19-G4 contact stands out as an exception in this regard as a Glu19Ala mutation to ΔK50ΔM51 and a G4A mutation to the RNA loop each had little effect on binding. This was the only mutation tested that showed indifference toward the smaller hairpin loops; no mutations increased affinity. Given that all other contacts are retained, this may be the result of only a slight shift in the RNA orientation on the protein scaffold, relative to the native interaction, occluding Glu19 from contact with the smaller hairpin loops.

Another significant result of this work is the first demonstration of ΔK50ΔM51 binding to a 7-nt loop hairpin and it does so with good affinity (K_D = 14.5 μM). This lowers the size limit of the loop that ΔK50ΔM51 can potentially accommodate in a potential RNA hairpin binding partner from 8-nt to 7-nt. This results in a higher number of potential targets in the pool of naturally occurring RNAs. ΔK50ΔM51-derivatives may in fact bind RNA hairpins with smaller loops, which we plan to explore in future studies.

The SAR determined here will be helpful to inform future engineering efforts centered around derivatives of the U1A RRM. Our results illuminate specific positions where ΔK50ΔM51 engages hairpin loops that are more comparable in size to most physiologically relevant hairpins. Of particular interest are contacts maintained by Asn15 and Asn16, which contribute to the sequence selectivity of the native interaction. These positions will likely be of elevated importance for redirecting the sequence specificity of the ΔK50ΔM51 scaffold. In aggregate, these results suggest that ΔK50ΔM51 is a suitable template scaffold for rational design or molecular evolution efforts to develop new recognition elements targeted toward

physiologically relevant RNA hairpins.

3.6 Methods

Molecular Cloning pET3d plasmids containing gene inserts that encode either U1A or ΔK50ΔM51 were generously provided by Professor Laird-Offringa. Mutants used in this work were generated using overlap PCR. Described below is an example protocol for introducing the F56A mutation into ΔK50ΔM51. The 5′- half of the ΔK50ΔM51 open reading frame was amplified using forward primer U1A FP NcoI (5′-CATGCCATGGCCCAGGTGCAG-3′) and reverse primer RP U1A F56A (5′-CCTCCTTGAAGATGACCGCGGCTTGGCCCCTCAGGCTCCG-3′) to yield a 211-nt amplicon. Similarly, the 3′- half of ΔK50ΔM51 was amplified using forward primer FP U1A F56A (5′-CGGAGCCTGAGGGGCCAAGCCGCGGTCATCTTCAAGGAGG-3′) and reverse primer U1A RP NotI (5′-ATAGTTTAGCGGCCGCAACCGG-3′) to yield a 203-nt amplicon. Both amplicons were individually analyzed and purified from a 1% agarose gel. They were then mixed together with 1X Vent polymerase reaction buffer, dNTPs, and Vent DNA polymerase (New England Biolabs). This mixture was then incubated for ten cycles of polymerase chain reaction (PCR), (94 °C for 10 seconds, 58 °C for 60 seconds, 72 °C for 60 seconds, and repeat for 10 cycles). The reaction was cooled to 4 °C so that forward primer U1A

FP NcoI (above) and reverse primer U1A RP NotI (above) could be added to the mixture. Following these additions, the reaction was allowed to proceed through 30 more cycles of PCR using the same times and temperatures. The desired 374-bp product encoding ΔK50ΔM51_F56A was analyzed and purified from a 1% agarose gel (agarose from Amaresco, 10X TBE buffer from Life Science Products inc, ethidium bromide from Sigma-Aldrich). A ΔK50ΔM51_F56A insert was then prepared for molecular cloning by digestion with restriction enzymes NcoI-HF and NotI-HF (both from New England Biolabs) in the appropriate reaction buffer, as recommended by the manufacturer, for two hours at 37 °C and then purified on a DNA affinity silica column (Qiagen). Similarly, pET3d plasmid was digested under the same conditions for three hours followed by addition of Calf Intestine Phosphatase (New England Biolabs) to the reaction and another one hour of incubation at 37 °C. Digestion of the DNA was verified on a 1% agarose gel and then cut vector was purified from the gel.

Ligation and Transformation Digested insert and digested vector were mixed at a 10 to 1 molar ratio with Quick Ligase (New England Biolabs) and 1X reaction buffer for ~20 minutes at RT, as recommended by the manufacturer. One mL of the crude ligation reaction was added to 25 mL of chemically competent 5-alpha E.coli (New England Biolabs) and incubated on ice for ~20 minutes. The mixture was then heat-shocked in a 42 ° C water bath for 30 seconds to facilitate DNA transformation and then allowed to incubate on ice for another 2 minutes. Cells were rescued with 950 mL of S.O.C. (supplied by the manufacturer) and incubated at 37 °C with shaking at 250 RPM for ~60 minutes. Cells were plated on pre-warmed LB-carbenecillin

plates (LB-millers broth from Fisher, Agar from Fisher, carbenecillin from GoldBio Technologies) and then plates were wrapped in parafilm (VWR) and incubated at 37 °C for 12-16 hours. Bacterial colonies growing on the plates were picked and diluted into 5 mL LB supplemented with 100 mg/mL cabenicillin and allowed to grow at 37 °C with shaking at 250 RPM for another 12-16 hours. Plasmid DNA was isolated from 5 mL cultures using a mini prep kit (Qiagen) as recommended by the manufacturer. The sequence of the newly created construct was verified using standard techniques performed by Genewiz Inc. ©. All mutants were prepared in a similar manner using the appropriate primers from the list below.

Protein Expression and Purification A construct encoding the appropriate mutant was transformed in to chemically competent BL-21 (DE3) E.coli as described above. Overnight cultures of the cells were inoculated into 2.5 L of LB media and incubated at 37 °C with shaking at 250 RPM until an OD_{600} of ~0.6 was reached (~3-4 hours). Expression of the desired mutant was induced by addition of IPTG (GoldBio Technologies) to a working concentration of 1 mM and incubated with shaking at 25 °C for at least four hours. Cells were harvested by centrifugation at 5000 RPM for ten minutes and then lysed by sonication. Lysate was centrifuged for 10 minutes at 15000 RPM and the supernatant was collected. The desired ΔK50ΔM51 variant was isolated from the soluble fraction by nickel affinity chromatography. Briefly, the supernatant was incubated with nickel-NTA resin (Fisher) with agitation at 4 °C for 10 minutes, then centrifuged for 10 minutes at 5000 RPM and the supernatant discarded. The resin was washed using an imidazole gradient as follows: 50 mL of 20 mM, 10 mL of 30 mM, 10

mL of 50 mM, and elution with 5 mL of 400 mM imidazole in HEPES buffer (Hepes-NaOH 10 mM, KCl 50 mM, $MgCl_2 \cdot 6H_2O$ 1mM, NaCl 30 mM, EDTA 1 mM). Eluted protein was dialyzed into HEPES without imidazole to remove significant traces of imidazole, yielding pure protein suitable for downstream applications. Protein concentration was determined using an extinction coefficient of 7450 L mol^{-1} cm^{-1}.

Affinity Measurements by Fluorescence Polarization The appropriate purified ΔK50ΔM51 variant (see above) was brought to a concentration of 1 mM and then serially diluted using a 1.7- fold dilution factor to generate a 24 sample dilution series. 5′-fluorescein labeled RNA oligonucleotides were ordered from Integrated DNA Technologies as RNase-free HPLC purified lyophilized pellets. The appropriate RNA oligo pellet was resuspended to a stock concentration of 10 μM. A master mix containing 10% NP-40, 40 nM 5¢-fluorescein labeled RNA oligo, and HEPES buffer (see above) was mixed. Prior to analysis, the mastermix was heated to 95 °C for two minutes and then plunged into ice to ensure hairpin formation. Mixing 20 mL of the ΔK50ΔM51 dilution series with 20 mL of the RNA mix gave 40 mL reactions containing 20 nM RNA and protein concentrations ranging from 500 mM to 0.002 mM. Fluorescence polarization measurements were made on a black 384-well polystyrene plate (Corning) at RT with a Perkin-Elmer Victor V multimode microplate reader. Data was processed using KaleidaGraph (Synergy Software) to determine RNA dissociation constants by fitting the data to single-site binding isotherm.

3.7 Proteins Used In This Work

U1A ΔK50ΔM51

nucleic acid sequence

ATGGCCCAGGTGCAGCTGCAGGTCGACATGGCAGTTCCCGAGACGCGTCCTAAC
CACACTATTTATATCAACAACCTCAATGAGAAGATCAAGAAGGATGAGCTCAAAA
AGTCCCTGTACGCCATCTTCTCCCAGTTTGGCCAGATCCTGGATATCCTGGTAT
CACGGAGCCTGAGGGGCCAAGCTTTTGTCATCTTCAAGGAGGTCTCGAGCGCCA
CCAACGCCCTGCGCTCCATGCAGGGTTACCCTTTCTATGACAAACCTATGCGTA
TCCAGTATGCGCGCACCGACTCAGATATCATTGCCAAGATGAAAGGCACCTTCG
GATCGGTCGACTCTAGAGGATCCCCGGTTGCGGCCGCACATCATCACCATCATC
ACGTGGCCGCAGAACAAAAACTCATCTCAGAAGAGGATCTGAATGGGGCCGCAT
AG

amino acid sequence

MAQVQLQVDMAVPETRPNHTIYINNLNEKIKKDELKKSLYAIFSQFGQILDILVSRSLRG
QAFVIFKEVSSATNALRSMQGYPFYDKPMRIQYARTDSDIIAKMKGTFGSVDSRGSPVA
AAHHHHHHHVAAEQKLISEEDLNGAA

3.8 Supplemental Data

Figure S3.1 Fluorescence polarization isotherms for U1A ΔK50ΔM51, and mutants thereof, binding to the 8-nt loop variant of U1hpII. Error is the standard deviation of four separate experiments.

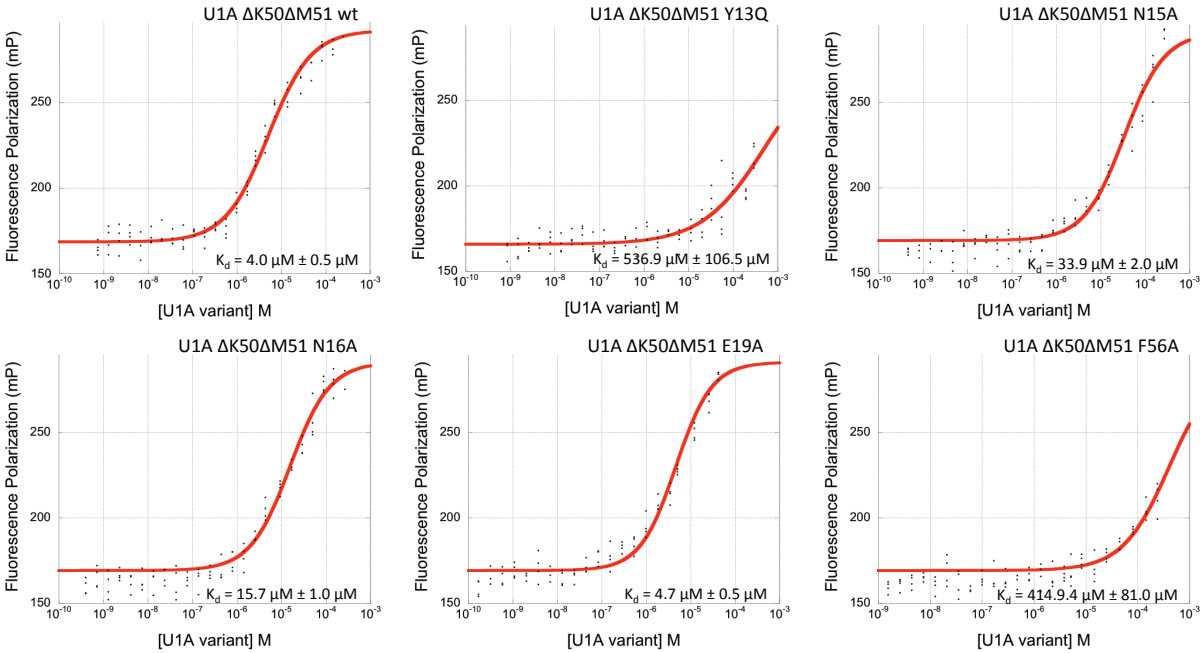


Figure S3.2 Fluorescence polarization isotherms for U1A ΔK50ΔM51, and mutants thereof, binding to the 7-nt loop variant of U1hpII. Error is the standard deviation of four separate experiments.

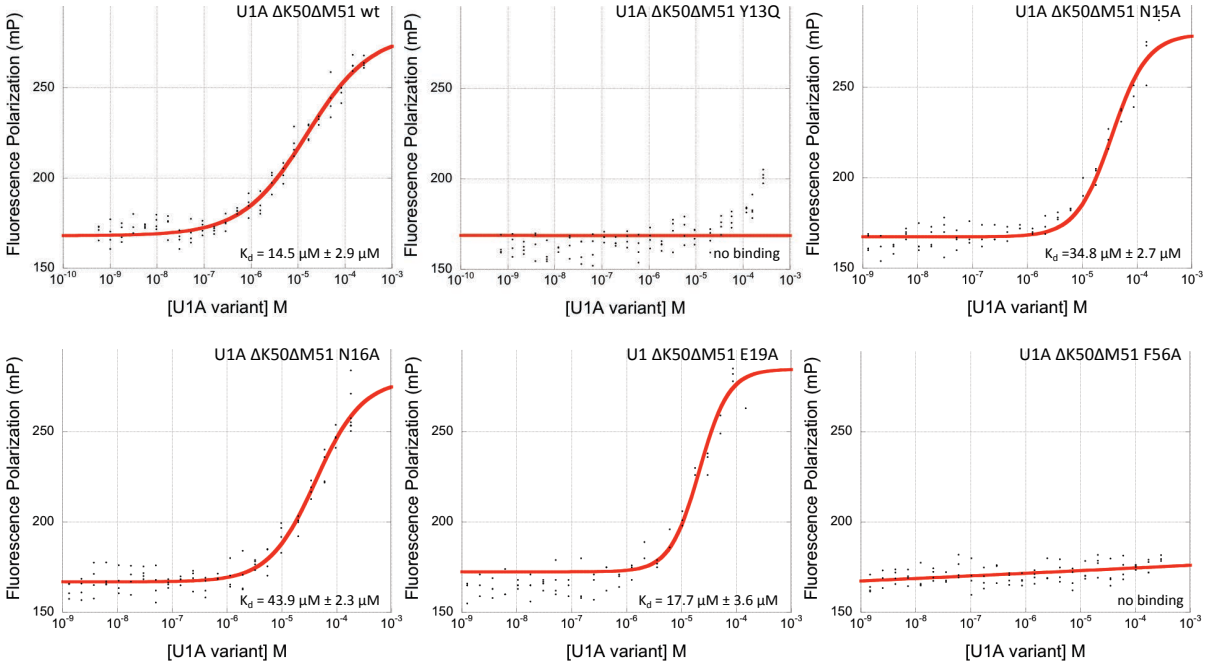
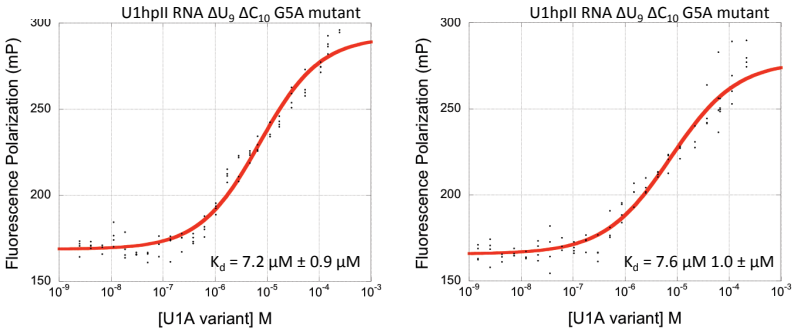


Figure S3.3 Fluorescence polarization isotherms for U1A ΔK50ΔM51binding to G4A mutants of both the 8- and 7-nt loop variants of U1hpII. Error is the standard deviation of four separate experiments.



REFERENCES

(1) Esteller, M. *Nat. Rev. Genet.* **2011**, *12*, 861.

(2) Osborne, R. J.; Thornton, C. a. *Hum. Mol. Genet.* **2006**, *15 Spec No*, R162.

(3) Cooper, T. a; Wan, L.; Dreyfuss, G. *Cell* **2009**, *136*, 777.

(4) Gallego, J.; Varani, G. *Acc Chem Res* **2001**, *34*, 836..

(5) Dervan, P. B.; White, S.; Szewczyk, J. W.; Turner, J. M.; Baird, E. E. *Nature* **1998**, *391*, 468.

(6) Cho, J.; Parks, M. E.; Dervan, P. B. *Proc. Natl. Acad. Sci.* **1994**, *92*, 10389.

(7) Dervan, P. B.; Edelson, B. S. *Curr. Opin. Struct. Biol.* **2003**, *13*, 284.

(8) Egholm, M.; Buchardt, O.; Nielsen, P. E. *J. Am. Chem. Soc.* **1992**, *114*, 1895.

(9) Nielsen, P. E. *Curr. Opin. Struct. Biol.* **1999**, *9*, 353.

(10) Nielsen, P. E. *Chem. Biodivers.* **2010**, *7*, 786.

(11) Disney, M. D.; Labuda, L. P.; Paul, D. J.; Poplawski, S. G.; Pushechnikov, A.; Tran, T.; Velagapudi, S. P.; Wu, M.; Childs-Disney, J. L. *J. Am. Chem. Soc.* **2008**, *130*, 11185.

(12) Paul, D. J.; Seedhouse, S. J.; Disney, M. D. *Nucleic Acids Res.* **2009**, *37*, 5894.

(13) Childs-Disney, J. L.; Yildirim, I.; Park, H.; Lohman, J. R.; Guan, L.; Tran, T.; Sarkar, P.; Schatz, G. C.; Disney, M. D. *ACS Chem. Biol.* **2014**, *9*, 538.

(14) Thomas, J. R.; Hergenrother, P. J. *Chem. Rev.* **2008**, *108*, 1171.

(15) Tor, Y. *ChemBioChem* **2003**, *4*, 998.

(16) Cléry, A.; Blatter, M.; Allain, F. H. T. *Curr. Opin. Struct. Biol.* **2008**.

(17) Chen, Y.; Varani, G. *FEBS J.* **2013**, *280*, 3734.

(18) Oubridge, C.; Ito, N.; Evans, P. R.; Teo, C. H.; Nagai, K. *Nature* **1994**, *372*, 432.

(19) Allain, F. H.; Howe, P. W.; Neuhaus, D.; Varani, G. *EMBO J.* **1997**, *16*, 5764.

(20) Katsamba, P. S.; Myszka, D. G.; Laird-Offringa, I. A. *J. Biol. Chem.* **2001**, *276*, 21476.

(21) Nolan, S. J.; Shiels, J. C.; Tuite, J. B.; Cecere, K. L.; Baranger, A. M. *J Am Chem Soc* **1999**, *121*, 8951.

(22) Law, M. J.; Chambers, E. J.; Katsamba, P. S.; Haworth, I. S.; Laird-Offringa, I. a. *Nucleic Acids Res.* **2005**, *33*, 2917.

(23) Shiels, J. C.; Tuite, J. B.; Nolan, S. J.; Baranger, A. M. *Nucleic Acids Res.* **2002**, *30*, 550.

(24) Showalter, S. a; Hall, K. B. *J. Mol. Biol.* **2004**, *335*, 465.

(25) Jessen, T. H.; Oubridge, C.; Teo, C. H.; Pritchard, C.; Nagai, K. *EMBO J.* **1991**, *10*, 3447.

(26) Yan Fan. *Thesis Diss.* **2009**.

(27) Allain, F. H.; Howe, P. W.; Neuhaus, D.; Varani, G. *EMBO J.* **1997**, *16*, 5764.

(28) Katsamba, P. S.; Bayramyan, M.; Haworth, I. S.; Myszka, D. G.; Laird-Offringa, I. a. *J. Biol. Chem.* **2002**, *277*, 33267.

(29) Kormos, B. L.; Pieniazek, S. N.; Beveridge, D. L.; Baranger, A. M. *Biopolymers* **2011**, *95*.

(30) Law, M. J.; Rice, A. J.; Lin, P.; Laird-Offringa, I. a. *RNA* **2006**, *12*, 1168.

(31) Kranz, J. K.; Hall, K. B. *J. Mol. Biol.* **1999**, *285*, 215.

(32) Kranz, J. K.; Hall, K. B. *J. Mol. Biol.* **1998**, *275*, 465.

CHAPTER FOUR

REDIRECTING THE BINDING SPECIFICITY OF AN RNA RECOGNITION MOTIF TOWARD HIV-1 TRANS-ACTING RESPONSE ELEMENT HAIRPIN RNA

4.1 Introduction

The growing exploration of RNA biochemistry has illuminated numerous biological processes that are controlled or regulated by coding and non-coding RNA molecules[1–5]. Functional RNAs may represent new entry points to manipulate living systems and correct human disease-states. These RNAs typically adopt a defined tertiary structure that informs their biological activity. Exacting control over RNA-dependent biochemical processes will first require recognition elements that tightly bind their folded RNA target in the complex cellular environment. Unfortunately, structured RNAs remain an elusive target for classical approaches to molecular recognition, owing largely to low functional group diversity, a complex electrostatic signature, and a highly dynamic shape. Developing reagents that accommodate these demanding properties will likely require an unconventional class of molecular probes.

We have previously presented an alternative approach to RNA recognition centered on semi-design of the U1A scaffold, an RNA Recognition Motif (RRM) protein borrowed from Nature[6]. In **Chapter 3**, we addressed concerns about the limited potential scope of RNA hairpin binding partners that the U1A scaffold may be able to accommodate. U1A's cognate RNA, U1hpII, contains an unusually large hairpin loop (10-nt)[7–9], however most physiological RNA hairpins have smaller loops (4-8 nt). Laird-Offringa and coworkers recently reported a

double deletion mutant, ΔK50ΔM51, that binds native U1hpII hairpin RNA and a truncated loop variant of U1hpII (8-nt loop) **Figure 3.3B** with similar affinities[10]. These binding properties are in contrast to wt U1A, suggesting that derivatization of ΔK50ΔM51, as opposed to wt U1A, may be a preferred strategy for targeting physiologically relevant RNA.

Here, we tackle another fundamental question regarding our approach toward semi-design of the U1A scaffold: can the RNA binding specificity of this scaffold be altered by mutating putative RNA binding regions? A beneficial feature of U1A is that its RNA binding interface is well annotated, which should aid in this interrogation. U1A engages its RNA binding partner through contacts on multiple regions, or sub-domains, of the protein. Predominantly, three of the most well studied sub-domains of the protein contribute to stabilize U1hpII binding. First, within the β-sheet face region, multiple researchers have shown that aromatic residues are critical for binding[8,11,12]. They π-stack in non-selective interactions with the aromatic groups of U1hpII loop nucleotides **Figure 4.1A** *red*, **Figure 3.2A**. Additionally, researchers in the Baranger lab have meticulously characterized many of the binding contacts on a second region, the β1-α1 loop **Figure 4.1A** *green*, **Figure 3.2B**, specifically highlighting contributions to the sequence selectivity of the interaction[13,14]. The third region, the β2-β3 loop, protrudes through the hairpin loop of U1hpII in the bound state, as seen in the crystal structure[7] **Figure 4.1A** *blue*, **Figure 3.2C**. This positioning strongly suggests that this sub-domain imparts a degree of shape complementarity for U1hpII, which is consistent with a number of studies focusing on the β2-β3 loop[10,15–18] (*see below*).

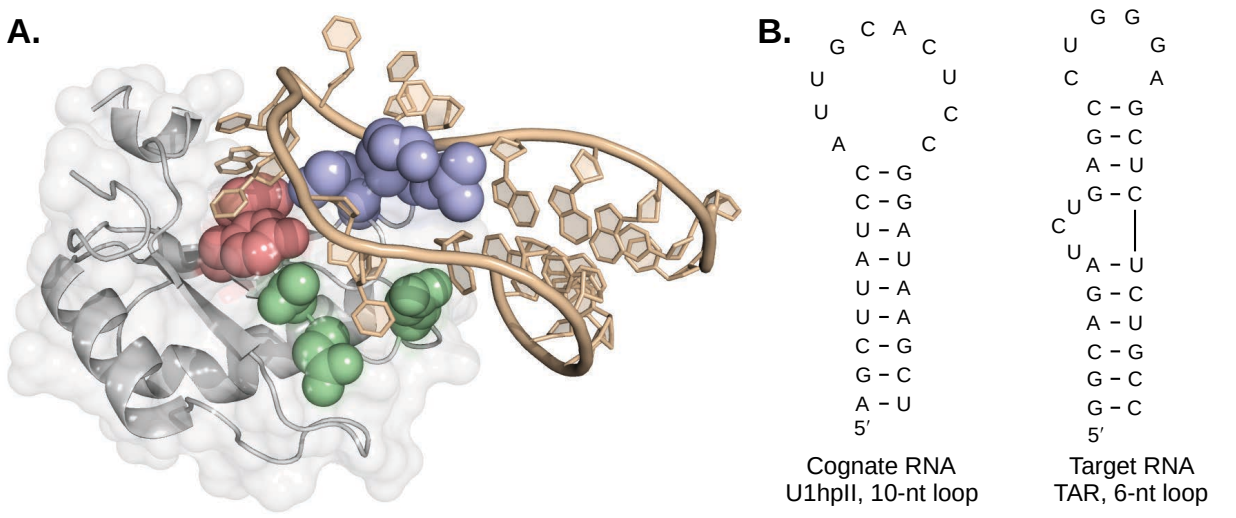


Figure 4.1 A Crystal structure of U1A in complex with its cognate RNA hairpin, U1hpII. Regions of interest to this study are highlighted as spheres, the β2-β3 loop in blue, the β1-α1 loop in green, and aromatic residues on the β-sheet face in red. **B** Nucleic acid sequence and hairpin secondary structures of RNAs relevant to this study. U1hpII, left, is the cognate binding partner of wt U1A and contains an unusually large hairpin loop, 10-nt. HIV-1 TAR RNA, right, is integral to HIV-1 infectivity. It contains a smaller hairpin loop, 6-nt, a stem loop bulge, and is not sequence homologous to U1hpII. (PDB code: 1URN)

The U1A protein scaffold displays a number of properties that suggest it may lend itself to serving as a general scaffold for folded RNA recognition. First, U1A's native RNA binding partner, U1hpII, adopts a defined tertiary structure in the shape of a hairpin **Figure 4.1B**. The high affinity of this interaction inside a cell raises the possibility that the overall structure of U1A is very well suited for nucleic acid-selective recognition of RNA hairpins. Acknowledging that U1A was evolved over millions of years to selectively bind it's RNA partner in a milieu of other cellular nucleic acids also suggests that RNA target selectivity may be an intrinsic feature of this scaffold. Second, the U1A scaffold demonstrates an inherent degree of plasticity, as previously reported mutations to surface residues generally do not disrupt the favorable stability or solubility attributes of the protein[6,10,15,18,19]. This is beneficial because redirecting the RNA binding specificity of the U1A scaffold will undoubtedly require modifications to surface

residues in order to engage new RNA binding partners.

Finally, and perhaps most compelling, multiple reports seem to suggest that directed mutations do in fact alter the RNA binding specificity of U1A. As mentioned previously, in **Chapter 3** we expand on a report by Laird-Offringa and coworkers that a double deletion to the β2-β3 loop allows the resulting mutant, ΔK50ΔM51, to bind the cognate RNA, U1hpII, and trunctated variants of U1hpII with similar affinity[106] **Figure 4.2A**. In another example, researchers identified mutations to the same region of U1A, the β2-β3 loop, that improve affinity for the cognate U1hpII RNA[18]. Additionally, Mattaj and coworkers also demonstrated the importance of the β2-β3 loop by grafting this region from a closely related RRM onto U1A[15] **Figure 4.2B**. This endowed the hybrid RRM with an increased affinity for an RNA hairpin with a similar size and sequence loop to U1hpII. Moreover, Baranger and coworkers reported that a mutation to the β1-α1 loop region, Glu19His, redirects the protein to selectively bind a sequence variant, G4U, of U1hpII[13] **Figure 4.2C**. Taken together, this limited data set suggests that changes to putative RNA binding regions on U1A (and by relation ΔK50ΔM51) are able to compensate for changes in the U1hpII binding partner and alter the binding selectivity.

A potential limitation to altering U1A, or ΔK50ΔM51, binding specificity is that in all examples listed above, the new RNA binding partner showed appreciable sequence and loop size homology with U1hpII RNA. In fact, no RRM has previously been reprogrammed to preferentially bind a non-homologous RNA binding partner, meaning the proposed strategy is largely untested. However, tangential evidence, including the analogous example of PUF and

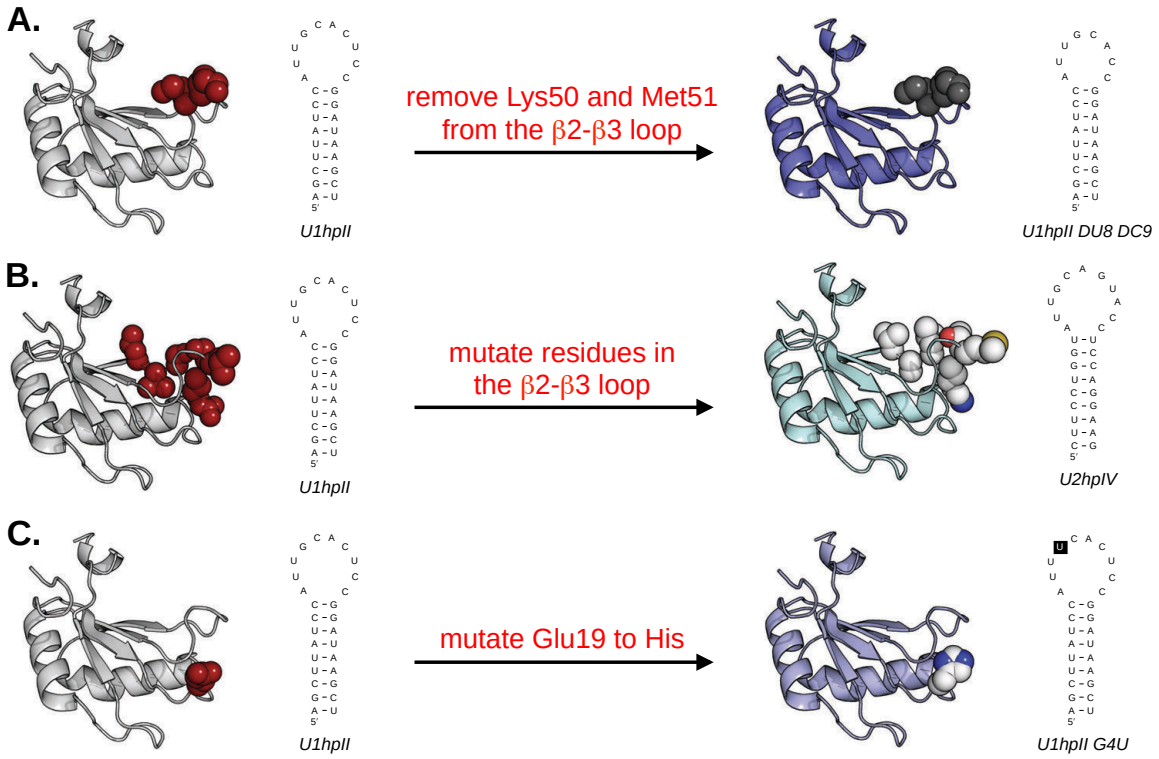


Figure 4.2 Through the characterization of the U1A-U1hpII binding interaction, researchers have indirectly shown that the specificity of U1A can be manipulated by mutations to key residues. **A** The ΔLys50ΔMet51 mutant of U1A binds both wild type and truncated variants of the U1hpII RNA hairpin, unlike wtU1A. **B** Mattaj and coworkers redirected the binding specificity of U1A to bind a homologous RNA hairpin, U2hpIV by grafting the β2-β3 loop from a similar RRM protein. **C** U1A does not bind the G4U mutant of U1hpII RNA with very good affinity, however affinity could be restored by making a single point mutation, Glu19His. (PDB code: 1URN)

PPR repeat proteins **Figure 1.16** and the exceptional RNA hairpin-binding properties of the U1A scaffold led us to question: Can we expand the scope of these previous observations and fully customize the U1A scaffold for binding new RNA hairpins with no sequence homology to U1hpII? Furthermore, would U1A-derived RRMs retain favorable binding properties of the parent scaffold, such as high affinity and specificity? Conceptually, the U1A scaffold will serve as an RNA binding canvas, in the sense that we plan to erase putative U1hpII binding elements and install alternative elements that recognize a new RNA target **Figure 4.3**.

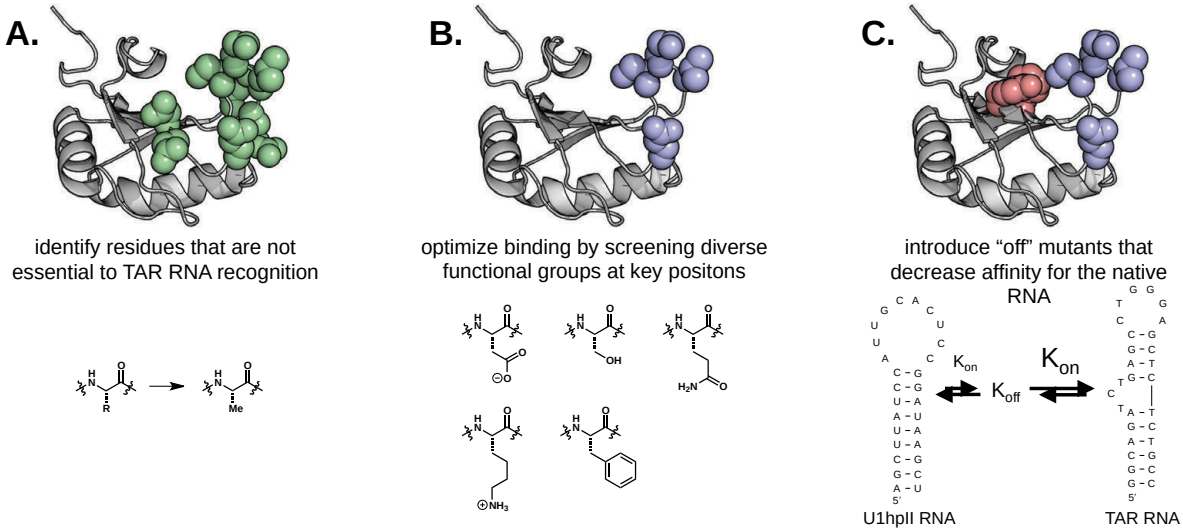


Figure 4.3 A Alanine scanning U1A at putative RNA binding residues (green) identifies disposable functional groups in the recognition of TAR RNA. **B** New functional groups are screened at key positions (blue) to identify new beneficial contacts with TAR RNA. **C** Mutations known to decrease binding to U1hpII, or “U1hpII off” mutations (red), are introduced into mutants that bind TAR RNA with good affinity. Disruption of native activity creates an orthogonally selective binder to the new RNA target. (PDB code: 1URN)

As an initial target to test our hypothesis, we chose the Trans Activating Response element (TAR) RNA hairpin, which plays a critical role in the infection cycle of HIV-1[20–23]. This hairpin RNA is an excellent initial target for three principle reasons. First, the hairpin loop of TAR is smaller than U1A’s cognate RNA, U1hpII, (6-nt instead of 10-nt) and does not share any sequence homology with U1hpII **Figure 4.1B**. Thus, this motif provides a unique RNA target to test the fundamental limits of reprogramming U1A, and by relation ΔK50ΔM51, specificity. Second, the TAR RNA hairpin has frequently been the target of other designed ligands of various sizes[24–28]. This enables a side-by-side comparison with other approaches in the field of RNA recognition. Third, assays for TAR-dependent HIV-1 activity are well reported in the literature[29–31], which affords us the opportunity to evaluate the activity of newly identified RRMs *in vivo*. Together, these make TAR RNA an excellent candidate for proof of concept studies

centered on developing a new platform for RNA hairpin targeted-synthetic RRMs.

4.2 U1A and ΔK50ΔM51 Display Good Affinity for TAR RNA Hairpin

Given a new RNA target, we looked to measure the affinity of both native U1A and ΔK50ΔM51 for the TAR RNA hairpin to broaden our potential discovery of new binding interactions. Also, this provides a direct comparison of the native U1A and ΔK50ΔM51 scaffolds in binding a smaller hairpin loop RNA (6-nt), allowing us to to test our hypothesis that ΔK50ΔM51 is a better starting template for these size loops. As a first approximation of how U1A and ΔK50ΔM51 interface with RNA hairpins that are non-sequence homologous to U1hpII, we used fluorescence polarization to measure the affinities of both scaffolds for the TAR RNA hairpin. We were very pleased to find both interactions gave dissociation constants (K_Ds) of < 20 μM **Table 4.1 entries 1 and 10**. These represent entirely new protein-RNA interactions with comparable affinities to recently reported synthetic small molecule-RNA interactions. These binding data suggest a degree of "shape" selectivity in recognizing the hairpin structure of TAR RNA. Based on these encouraging results, we looked to further interrogate these new interactions and potentially identify mutations that improve their affinity.

Table 4.1 Alanine scanning of putative RNA binding residues on both the U1A and ΔK50ΔM51 protein scaffolds with their respective affinities for TAR RNA.

entry	protein scaffold	mutation[a]	nucleic acid target	K_D (μM)[b]	ΔG (kcal/mol)[c]	ΔΔG (kcal/mol)[d]
1	U1A	-	TAR RNA	19.2 (±4.9)	-6.43	-
2	U1A	Asn15Ala	TAR RNA	11.7 (±0.9)	-6.72	-0.29
3	U1A	Asn16Ala	TAR RNA	8.8 (±1.4)	-6.89	-0.46
4	U1A	Glu19Ala	TAR RNA	5.9 (±0.8)	-7.13	-0.70
5	U1A	Ser46Ala	TAR RNA	15.1 (±2.6)	-6.57	-0.14
6	U1A	Ser48Ala	TAR RNA	14.2 (±1.3)	-6.61	-0.18
7	U1A	Leu49Ala	TAR RNA	5.2 (±0.5)	-7.20	-0.77
8	U1A	Lys50Ala	TAR RNA	39.6 (±12.2)	-6.00	+0.43
9	U1A	Met51Ala	TAR RNA	4.9 (±0.6)	-7.24	-0.81
10	ΔK50ΔM51	-	TAR RNA	15.9 (±2.3)	-6.54	-
11	ΔK50ΔM51	Asn15Ala	TAR RNA	91.1 (±8.8)	-5.51	+1.03
12	ΔK50ΔM51	Asn16Ala	TAR RNA	62.1 (±5.4)	-5.74	+0.80
13	ΔK50ΔM51	Glu19Ala	TAR RNA	10.2 (±1.1)	-6.80	-0.26
14	ΔK50ΔM51	Ser46Ala	TAR RNA	24.5 (±6.0)	-6.29	+0.25
15	ΔK50ΔM51	Ser48Ala	TAR RNA	33.9 (±16.0)	-6.09	+0.45
16	ΔK50ΔM51	Leu49Ala	TAR RNA	91.6 (±39.7)	-5.50	+1.04

[a]All point mutations are derived from either U1A (entries 2-9) or ΔK50ΔM51 (entries 11-16). [b]Reported error for each measurement represents the standard deviation of at least three experiments. [c]ΔG is the free energy of the protein in complex with TAR RNA calculated with the equation $\Delta G = -RT \ln K_D$. [d]ΔΔG is the difference in binding energy between the complexe in entry 1 (for U1A-derived mutants) or in entry 10 (for ΔK50ΔM51-derived mutants) and the indicated mutant thereof. Experiments were performed in Hepes buffer, pH 7.4 at 25 °C (see **Methods and Materials**).

4.3 Alanine Scanning of Putative RNA Binding Residues Reveals Important Interactions in the TAR RNA Interface

To gain further insight into the mechanism of recognition for both of these newly identified interactions, we performed alanine scanning on putative RNA binding regions of the proteins **Figure 4.3A**. This would allow us to identify regions on the protein that are non-essential or deleterious to engaging TAR RNA. Given the demonstrated importance of the β2-β3 loop[6,10,15,18,32] **Figure 4.1A** *blue* (detailed above and in section **3.2**), we included residues from this region into the analysis, specifically: Ser46, Ser48, Leu49, Lys50, and Met51 **Figure 4.4A**. Previous research has shown that removing Arg47, or analogous residues in other RRMs, is generally deleterious to RNA affinity[19,32–35], so we elected to exclude this position. Alanine mutations at position 46 and 48 showed modest increases in affinity **Table 4.1 entries 5 and**

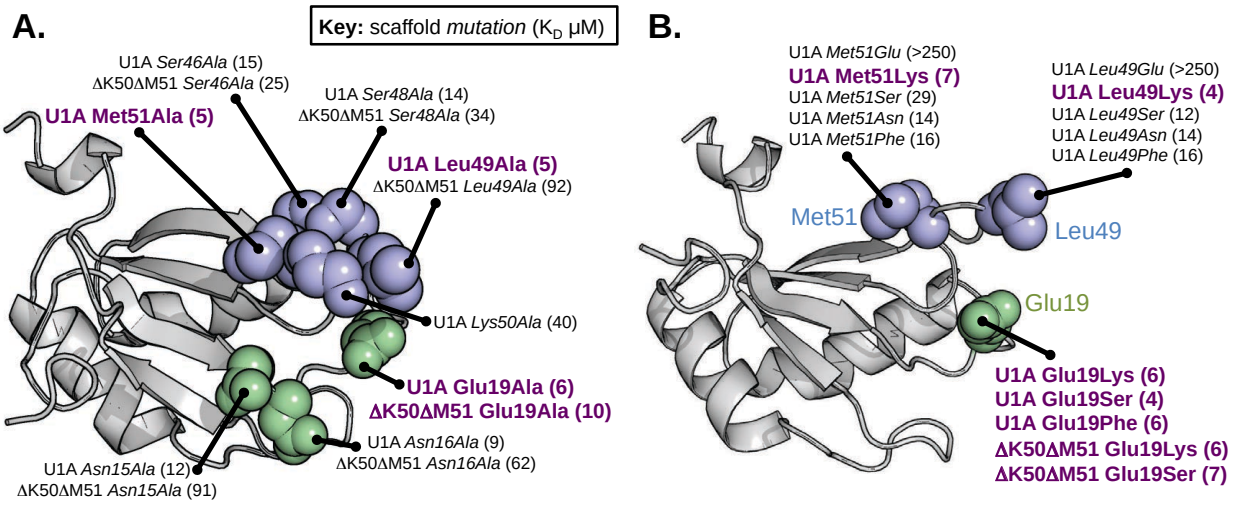


Figure 4.4 A Map of the U1A scaffold representing effects of alanine mutations at putative binding residues. Resultant K_D's are shown in parentheses. Mutations that gave the best affinities are bolded in purple. **B** Key positions on the scaffolds, identified from alanine scanning in **A**, were functionally diversified by introducing new residues at these positions. Resultant K_D's are shown in parentheses. Mutations that gave the best affinities are bolded in purple. An exhaustive list of varied functional group mutations with associated K_D's is provided in **Table 4.2**. (PDB code: 1URN)

6, while mutations to position 49 and 51 gave much greater increases of ~3.7- and ~3.9-fold, respectively **Table 4.1 entries 7 and 9**. Residues in the β1-α1 loop engage in hydrogen bonding interactions that are closely tied to the native affinity and selectivity of U1A[13,14,36] **Figure 4.1A**, **Figure3.2B.** All three alanine substitutions made to important residues on the β1-α1 loop, Asn15, Asn16, and Glu19 **Figure 4.4A**, increased the affinity of the interaction **Table 4.1 entries 2,3, and 4**. Glu19Ala showed the greatest increase, ~3.2- fold **Table 4.1 entry 4**. The increase in binding upon, essentially, removing the functional group at select positions suggests that these functional groups are not optimal for recognizing TAR RNA, and may in fact be deleterious in this capacity.

Next, we performed the same alanine scan on the ΔK50ΔM51 scaffold to discern how key residues of this scaffold interface with TAR RNA. In contrast to the U1A scaffold, all but one of

the alanine mutants showed a decrease in affinity relative to the parent ΔK50ΔM51 scaffold **Table 4.1 entries 11-16**. The ΔK50ΔM51_Glu19Ala mutant was the exception, showing a ~1.6-fold increase in affinity **Table 4.1 entry 13**. Contrary to our initial assessment, these results seem to indicate that the wtU1A scaffold may be a better initial starting point to generate new TAR RNA-protein ligands.

From this initial exercise, we identified three variants of the U1A scaffold that demonstrated a 3.2-fold or greater increase in affinity over the wild type: U1A_Glu19Ala, U1A_Leu49Ala, and U1A_Met51Ala **Table 4.1 entries 4, 7, and 9**. We selected these three U1A variants for further analysis. ΔK50ΔM51_Glu19Ala was not among the tightest binders of TAR RNA **Table 4.1 entry 13**, however, we elected to carry this variant forward for further analysis since this beneficial mutation was common between the two scaffolds. The alanine scanning data suggest that residues at position 19 in both U1A and ΔK50ΔM51 scaffolds, and positions 49 and 51 in the U1A scaffold play an important role in binding TAR RNA.

4.4 New Chemical Functionality at Key Positions Improves Affinity for the TAR RNA hairpin

Taking what we learned from the alanine scanning data, we next looked to further interrogate the newly identified key positions on the U1A and ΔK50ΔM51 scaffolds. We introduced a diverse but focused library of unique functional groups at each position. By testing

the functional group tolerance at these positions, we hope to potentially identify residues that facilitate new productive contacts with the bound TAR RNA, thus improving the affinity of the interaction. Specifically, we included either a negatively charged (glutamic acid), positively charged (lysine), hydrogen bond donating (serine), amide (glutamine), or hydrophobic aromatic (phenylalanine) functionality **Figure 4.4B**, **Table 4.2**.

Installing the negatively charged glutamic acid at either position 49 or 51 was catastrophic to TAR RNA binding affinity as fluorescence polarization isotherms for both interactions failed to saturate at concentrations up to 500 μM protein **Table 4.2 entries 5 and 10**, see **Supplemental Data: Figure S2**. Conversely, we saw a universal improvement in affinity when incorporating a positively charged lysine residue, when compared to the wild type scaffolds, as all of these variants gave ~10^{-6} K_Ds **Table 4.2 entries 1, 6, 11, and 15**. Free amines, and other positively charged functional groups, are common in small molecules developed to bind RNA[37–41], likely due to non-specific charge attraction with the anionic phosphate backbone. Thus we remained cognizant that the observed TAR RNA binding affinities for these mutants may be non-specific and would require further evaluation.

Serine scanning at each key position gave a more diverse affinity profile. The U1A_Glu19Ser showed a ~4.7-fold improvement over the wild type scaffold **Table 4.2 entry 2**, while Met51Ser showed a ~1.6-fold decrease in affinity **Table 4.2 entry 12**. Installing an amide functionality showed an interesting contrast to the glutamic acid-substituted variants as asparagine was well tolerated at all key positions **Table 4.2 entries 3, 8, 13, and 17**. Phenylalanine substitutions also seemed to be well tolerated throughout all key positions; only

Table 4.2 Functional group diversity at key positions on both the U1A and ΔK50ΔM51 protein scaffolds with their respective affinities for TAR RNA.

entry	protein scaffold	mutation[a]	nucleic acid target	K_D (μM)[b]	ΔG (kcal/mol)[c]	ΔΔG (kcal/mol)[d]
1	U1A	Glu19Lys	TAR RNA	5.9 (±0.4)	-7.13	-0.70
2	U1A	Glu19Ser	TAR RNA	4.1 (±0.3)	-7.34	-0.91
3	U1A	Glu19Asn	TAR RNA	12.0 (±0.9)	-6.71	-0.28
4	U1A	Glu19Phe	TAR RNA	6.2 (±0.8)	-7.10	-0.67
5	U1A	Leu49Glu	TAR RNA	>250		
6	U1A	Leu49Lys	TAR RNA	3.9 (±0.4)	-7.37	-0.94
7	U1A	Leu49Ser	TAR RNA	12.4 (±2.3)	-6.69	-0.26
8	U1A	Leu49Asn	TAR RNA	14.4 (±2.1)	-6.60	-0.17
9	U1A	Leu49Phe	TAR RNA	24.1 (±4.3)	-6.30	+0.13
10	U1A	Met51Glu	TAR RNA	>250		
11	U1A	Met51Lys	TAR RNA	7.7 (±0.5)	-6.97	-0.54
12	U1A	Met51Ser	TAR RNA	29.9 (±5.7)	-6.17	+0.26
13	U1A	Met51Asn	TAR RNA	13.8 (±1.6)	-6.62	-0.19
14	U1A	Met51Phe	TAR RNA	16.4 (±2.3)	-6.52	-0.09
15	ΔK50ΔM51	Glu19Lys	TAR RNA	6.3 (±1.0)	-7.09	-0.55
16	ΔK50ΔM51	Glu19Ser	TAR RNA	7.0 (±1.9)	-7.03	-0.47
17	ΔK50ΔM51	Glu19Asn	TAR RNA	27.8 (±3.2)	-6.21	+0.33
18	ΔK50ΔM51	Glu19Phe	TAR RNA	12.6 (±1.8)	-6.68	-0.14

[a]All point mutations are derived from either U1A (entries 1-14) or ΔK50ΔM51 (entries 15-18). [b]Reported error for each measurement represents the standard deviation of at least three experiments. [c]ΔG is the free energy of the protein in complex with TAR RNA calculated with the equation $\Delta G = -RT \ln K_D$. [d]ΔΔG is the difference in binding energy between the complex in Table 1 entry 1 (for U1A-derived mutants) or in Table 1 entry 10 (for ΔK50ΔM51-derived mutants) and the indicated mutant thereof. Experiments were performed in Hepes buffer, pH 7.4 at 25 °C (see **Methods and Materials**).

U1A_Leu49Phe showed little to no decrease in affinity relative to the wild type scaffold **Table 4.2 entry 9**.

From these data we identified two mutants with improved affinity for TAR RNA relative to the best performing alanine scanning mutant, U1A_Met51Ala **Table 4.1 entry 9**. U1A_Glu19Ser and U1A_Leu49Lys each bound TAR RNA with Kds of ~4.1 μM and ~3.9 μM, respectively **Table 4.2 entries 2 and 6**. The best performing mutants now gave a combined population of ten U1A and ΔK50ΔM51 variants that bound the target RNA with a K_D of 7.7 μM or lower. These are remarkable data, as they indicate that mutations to the U1A and ΔK50ΔM51 scaffolds can improve their affinity for a new target RNA that is non-homologous to the U1A cognate RNA, U1hpII. This finding is consistent with our hypothesis and has significant

implications toward our larger goals of defining a general platform for folded RNA recognition.

4.5 Multiple Mutations Do Not Synergistically Improve Affinity for the TAR RNA hairpin

In the pursuit of engineering new protein function, it is common for multiple beneficial mutations to endow a synergistic effect to increase the desired activity[42–44]. To look for synergism, we generated another combinatorial library of U1A variants each containing two or three beneficial mutations identified from the previous data sets. Unfortunately, none of these variants demonstrated appreciable improvements in fitness **Supplemental Data: Figure S6**.While it is not the focus of this study to exhaustively characterize the structural orientation of these interactions, one may speculate that TAR RNA adopts different spatial positions when binding to different variants of the U1A scaffold. This would render a second beneficial mutation irrelevant in an alternate binding mode. In the absence of structural data on these novel interactions to guide informed placement of new mutations, we abandon this low throughput exploration of U1A sequence space.

By building on an inherent baseline affinity for the TAR RNA hairpin, we have shown that we can not only improve the strength of these new binding interactions but we are achieving very good overall affinities relative to recently reported RNA-binding small molecules. This exploration of RRM sequence space was far from exhaustive (32 variants analyzed). However, this data set does provide compelling evidence in support of our approach to generate new RNA

binding molecules and we saw fit to leave further optimization to a more high-throughput screening technique (Chapter 5).

4.6 RRM Variants Preferentially Bind RNA Over DNA

Achieving target selectivity is a major hurdle for RNA binding reagents. This attribute is particularly important for use *in vivo* or in conditions that mimic a complex cellular environment. In this context, preferential recognition of RNA over DNA is a relevant as well as a difficult task. The two nucleic acids share many structural and chemical features, complicating selective recognition efforts by small molecule reagents. Given the substantially greater structural complexity of our synthetic RRMs, we were therefore interested in assessing their nucleic acid selectivity.

We first measured the affinity of unmodified U1A and ΔK50ΔM51 for a DNA analogue of the TAR hairpin, in which all of the ribonucleotides were exchanged for 2′-deoxyribonucleotides and uracil replaced with thymine **Figure 4.5**. Amazingly, both scaffolds showed large decreases in their affinity for the TAR DNA analogue compared to TAR RNA, 7.6- and >30-fold decrease respectively **Table 4.3 entries 1 and 2**. This was very encouraging as these results strengthen our hypothesis that an inherent virtue of the parent RRM scaffold allows selective recognition of RNA. Also of significance, the unmodified scaffolds largely out perform other synthetic RNA-binding reagents in terms of nucleic acid selectivity[45,46,47].

Because all of our synthetic RRM variants were derived from these same RNA-selective

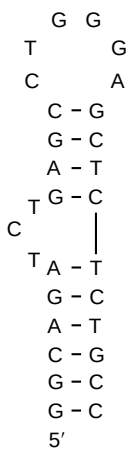


Figure 4.5 Nucleic acid sequence and secondary structure of the TAR DNA analogue. All of the the ribonucleotides were exchanged for 2′-deoxyribonucleotides and uracil replaced with thymine.

parent scaffolds, we expected that our variants would also be able to discriminate between the two chemically similar nucleic acids. We tested our ten best variants, taken from both the alanine scan and the functional group scan, for TAR DNA affinity. Mutants with a newly introduced lysine residue showed poor nucleic selectivity **Table 4.3 entries 3 and 5-7**. As we suspected, the increase in positive charge density may elicit non-specific electrostatic interactions with the poly-cationic phosphate backbone of both nucleic acids. Also, none of the ΔK50ΔM51-derived variants showed good nucleic acid selectivity **Table 4.3 entries 3 and 4**. However, five of the U1A-derived RRMs demonstrated at least 17-fold selectivity for the TAR RNA hairpin over the TAR DNA hairpin **Table 4.3 entries 8-12**. These data are impressive, given the chemical and structural similarities between TAR RNA and TAR DNA. This further suggests that desirable features of the U1A scaffold, in this case nucleic acid selectivity, are retained in variants with surface mutations. We carried these five RNA-selective U1A-derivatives forward for further analysis.

Table 4.3 Nucleic acid selectivity of U1A and ΔK50ΔM51, and variants thereof

entry	protein scaffold	mutation[a]	nucleic acid target	K_D (μM)[b]	fold selectivity[c]
1	U1A	-	TAR DNA	145.5 (±46.0)	7.6
2	ΔK50ΔM51	-	TAR DNA	>250	>15.7
3	ΔK50ΔM51	Glu19Lys	TAR DNA	29.8 (±5.8)	4.7
4	ΔK50ΔM51	Glu19Ser	TAR DNA	59.9 (±14.2)	8.6
5	U1A	Glu19Lys	TAR DNA	12.3 (±2.0)	2.1
6	U1A	Leu49Lys	TAR DNA	19.5 (±2.9)	5.0
7	U1A	Met51Lys	TAR DNA	54.3 (±7.9)	7.1
8	U1A	Glu19Ala	TAR DNA	100.1 (±33.4)	17.0
9	U1A	Glu19Ser	TAR DNA	91.9 (±33.4)	22.4
10	U1A	Glu19Phe	TAR DNA	156.4 (±80.0)	25.2
11	U1A	Leu49Ala	TAR DNA	88.2 (±22.0)	17.0
12	U1A	Met51Ala	TAR DNA	178.7 (±37.5)	36.5
13	U1A	-	U1hpII RNA	3.4 (±0.4) x 10^{-11}	<1
14	U1A	Glu19Ser	U1hpII RNA	11.4 (±1.6) x 10^{-6}	2.8
15	U1A	Glu19Phe	U1hpII RNA	36.1 (±6.8) x 10^{-6}	5.8
16	U1A	Glu19Ala	U1hpII RNA	2.6 (±0.2) x 10^{-6} [d]	<1
17	U1A	Leu49Ala	U1hpII RNA	<1 x 10^{-8}	<1
18	U1A	Met51Ala	U1hpII RNA	2.3 (±0.2) x 10^{-10} [d]	<1

[a]All point mutations are derived from either U1A or ΔK50ΔM51. [b]Reported error for each measurement represents the standard deviation of at least three experiments. [c]Fold-selectivity for TAR RNA was calculated by (K_D off-target nucleic acid/ K_D TAR RNA). [d]The binding affinity for these interactions has previously been reported, as shown in references X and Y, respectively. Experiments were performed in Hepes buffer, pH 7.4 at 25 °C (see **Methods and Materials**).

4.7 Synthetic RRMs Demonstrate Selectivity for TAR RNA Over the U1A Cognate RNA

In addition to DNA, other RNA molecules also contaminate *in vivo* or mock-*in vivo* experimental conditions. Target discrimination among other RNAs is essential for applied RNA-binding reagents. To address this property within the newly identified RRMs, we first looked to determine their affinity for the cognate RNA of the parent scaffold, the U1hpII hairpin RNA. Given the incredibly high affinity of the native interaction (K_D ~10^{-11}), this is a fitting test of selectivity as most sub-domains that engage U1hpII in the native complex are retained in these variants. Two of these binding interactions had been previously reported in the literature; Glu19Ala and Met51Ala both bind to U1hpII with low to sub-nanomolar affinity **Table 4.3 entries 16 and 18**. We determined the affinity of the three remaining RRMs for U1hpII by fluorescence polarization. Two of these RRMs, U1A_Glu19Phe and U1A_Glu19Ser, showed a

dramatic decrease in affinity for U1hpII, ~10^6-fold increase in K_D **Table 4.3 entries 14 and 15**. Amazingly, both of these synthetic RRMs bind TAR RNA with modest selectivity over U1hpII, 5.8- and 2.8-fold, respectively. The regions we chose to modify on U1A are inherently pivotal for U1hpII recognition. These mutations therefore logically translate to a drastic decrease in U1hpII affinity, concomitant with the more surprising increase in TAR RNA affinity. This approach has enabled the selectivity we observe for U1A_Glu19Phe and U1A_Glu19Ser.

4.8 Directed Mutations Disrupt Native Binding Function and Engender Greater Specificity

The rich literature precedent on the U1A-U1hpII binding interaction provides us a robust understanding of the associated binding mode. Inspired by the target selectivity observed above, we next looked to apply this resource to actively engineer "U1hpII off" mutations into our synthetic RRMs that would decrease their affinity for U1hpII **Figure 4.3c**. Ideally, carefully chosen "U1hpII off" mutations should not disrupt TAR RNA binding affinity, and thus should afford a new collection of RRMs with selectivity for the new RNA target. This exercise may also reveal new information about the TAR RNA binding mode.

We selected four point mutations that have been shown to greatly destabilize the native binding function of U1A. Introducing either a Tyr13Gln or Phe56Ala mutation to U1A greatly destabilizes the native interaction with U1hpII[8,32,33]. An Asn15Val mutation is also deleterious to U1hpII affinity[15,33]. From alanine scanning, we have shown that removal of the amide functional

group at this position slightly increased affinity for TAR RNA. Similarly, Glu19Ala has been shown to destabilize the native interaction by ~200-fold[13,14], while our data shows a ~3.2-fold increase in affinity for TAR RNA **Table 4.1 entry 4**.

We selected these four mutations, Tyr13Gln, Phe56Ala, Asn15Val, and Glu19Ala, as our "U1hpII off" mutations and cross-referenced them with the five selected TAR RNA binding U1A variants, U1A_Glu19Ala, U1A_Glu19Phe, U1A_Glu19Ser, U1A_Leu49Ala, and U1A_Met51Ala. This generated a library of 17 U1A double mutants **Table 4.4**.

The primary concern was that "U1hpII off" mutations would have an effect on TAR RNA binding affinity. To address this, we individually screened these double mutant RRMs to identify combinations that retain good affinity for TAR RNA. Many of the variants saw a significant decrease in TAR RNA binding affinity upon additional of "U1hpII off" mutations. However, five variants did retain < 15.0 μM K_D (as low as 6.9 μM) in complex with TAR RNA **Table 4.4 entries 1, 2, 6, 7, and 9**, suggesting these differences in the TAR RNA interfaces allow these RRMs to accommodate their respective "U1hpII off" mutation with little disturbance to TAR RNA affinity.

Next, to discern the selectivity of these five double mutants, we determined the U1hpII affinity for each. As expected, the "U1hpII off" mutations led to large drops in U1hpII affinity, ~10^6-fold. This generated three variants with appreciable selectivity for TAR RNA over U1hpII **Table 4.4 entries 1, 6, and 9**. These variants required only two rationally directed mutations to reprogram the RNA binding specificity of the scaffold. This nicely showcases the malleability of U1A binding specificity and encourages the further discovery of unique RNA

Table 4.4 Selectivity of "U1hpII off" mutants for TAR RNA over U1hpII RNA

entry	protein scaffold	mutation[a]	U1hpII off mutation[a]	TAR K_D (μM)[b]	U1hpII K_D (μM)[b]	fold selectivity[c]
1	U1A	Glu19Ala	Asn15Val	14.2 (±1.4)	48.4 (±12.4)	3.4
2	U1A	Glu19Ser	Asn15Val	11.3 (±0.8)	15.1 (±2.0)	1.3
3	U1A	Glu19Phe	Asn15Val	19.0 (±2.5)	ND	-
4	U1A	Leu49Ala	Asn15Val	39.2 (±9.6)	ND	-
5	U1A	Met51Ala	Asn15Val	46.0 (±13.4)	ND	-
6	U1A	Leu49Ala	Glu19Ala	6.9 (±0.8)	11.7 (±2.1)	1.7
7	U1A	Met51Ala	Glu19Ala	14.9 (±2.5)	15.8 (±2.7)	1.1
8	U1A	Glu19Ala	Tyr13Gln	21.7 (±5.3)	ND	-
9	U1A	Glu19Ser	Tyr13Gln	10.3 (±2.8)	63.6 (±8.9)	6.2
10	U1A	Glu19Phe	Tyr13Gln	58.0 (±24.1)	ND	-
11	U1A	Leu49Ala	Tyr13Gln	>250	ND	-
12	U1A	Met51Ala	Tyr13Gln	209.8 (±84.3)	ND	-
13	U1A	Glu19Ala	Phe56Ala	29.2 (±6.3)	ND	-
14	U1A	Glu19Ser	Phe56Ala	19.2 (±1.4)	ND	-
15	U1A	Glu19Phe	Phe56Ala	20.9 (±3.3)	ND	-
16	U1A	Leu49Ala	Phe56Ala	123.2 (±38.7)	ND	-
17	U1A	Met51Ala	Phe56Ala	176.9 (±98.2)	ND	-

[a]All point mutations are derived from U1A. [b]Reported error for each measurement represents the standard deviation of at least three experiments. [c]Fold-selectivity for TAR RNA was calculated by (K_D off-target nucleic acid/ K_D TAR RNA). Experiments were performed in Hepes buffer, pH 7.4 at 25 °C (see **Methods and Materials**).

binding reagents derived from this scaffold.

4.9 Synthetic RRMs Recognize TAR RNA In a Complex Mixture of Non-Specific RNAs

As a final indication of selectivity, it is also important that TAR RNA-binding RRMs identify their target in a complex solution of non-specific RNA molecules. Three synthetic RRMs were selected based on their desirable affinity and selectivity profiles. Their affinity for TAR RNA was measured in the presence of 10 molar excess tRNAs. These serve as an adequate representation of non-specific folded RNAs as they contain many secondary structure elements, including hairpins. It was therefore very gratifying to find that select mutants were minimally affected, if at all, by a superstoichiometric ratio of nonspecific competing RNAs **Table 4.5**. This

Table 4.5 Affinity of select mutants in the presence of 10 molar excess tRNAs

entry	protein scaffold	mutation[a]	U1hpII off mutation[a]	TAR K_D (μM)[b]	w tRNAs K_D (μM)[b]	fold selectivity[c]
1	U1A	Glu19Ser	-	4.1 (±0.3)	6.4 (±0.5)	1.6
2	U1A	Glu19Phe	-	6.3 (±0.8)	19.9 (±1.8)	3.2
3	U1A	Glu19Ser	Tyr13Gln	10.3 (±1.4)	10.2 (±0.7)	none

[a]All point mutations are derived from U1A. [b]Reported error for each measurement represents the standard deviation of at least three experiments. [c]Fold-selectivity for TAR RNA was calculated by (K_D TAR RNA in the presence of 10 molar excess tRNAs/ K_D TAR RNA). Experiments were performed in Hepes buffer, pH 7.4 at 25 °C (see **Methods and Materials**).

is an excellent demonstration of the highly precise nature of the newly identified protein-RNA interactions. Collectively, these data indicate that the newly identified synthetic RRMs possess an exceptional selectivity profile in comparison to most synthetic RNA binding reagents.

4.10 Folding and Secondary Structures are Preserved in U1A Variants

A large body of evidence suggests that U1A tolerates mutations to surface exposed residues without compromising the overall folded structure of the protein[6,18,19,33]. To ensure that previously unreported TAR RNA binding mutants retained secondary structure features of the parent scaffold, circular dichroism measurements were individually taken on U1A_Glu19Ser, U1A_Glu19Phe, and U1A_Glu19Ser:Tyr13Gln. When compared to circular dichromism scans of wild type U1A, these variants gave identical signatures **Supplemental Data: Figure S7**. This suggests that the observed binding properties are not derived from conformational changes in the protein but rather from improved contacts at the binding interface made by the new mutations.

4.11 Conclusions

Here we propose a new approach to solving the folded RNA recognition problem by redirecting an RRM scaffold from Nature to bind a different RNA hairpin target. A major obstacle to this approach is the viability of reprogramming the RNA binding selectivity this RRM scaffold while retaining high affinity and selectivity characteristics in the new interaction. Here, we have demonstrated that thoughtfully placed mutations on the U1A scaffold can improve the affinity for an orthogonal RNA hairpin target. The newly identified synthetic RRMs also have exquisite selectivity for the new target. These data strongly advocate for this as a viable strategy for generating new RNA hairpin binding reagents that identify their target in a biologically relevant context.

Through focused screening steps we have distilled down multiple synthetic RRMs with enhanced recognition properties for the HIV-1 TAR RNA hairpin. Alanine scanning of putative RNA binding regions on U1A identified positions with disposable functional groups in the new TAR RNA binding interface. Scanning a chemically diverse set of amino acids at select positions demonstrated that new chemical functionality on the scaffold facilitates improved binding contacts with the new target. RRMs with the tightest binding affinity recognize TAR RNA with dissociation constants of < 5 μM.

The newly reported synthetic RRMs outperform the majority of state of the art RNA binding small molecules with respect to target selectivity[26,39,48]. Most variants tested here show ≥ 17.0-fold selectivity for TAR RNA over a DNA analogue. This is likely afforded by the inherent RNA recognition properties of the parent RRM scaffold. By the dictates of our approach, the

initial screens generated U1A derivatives that contain mutations at putative U1hpII binding regions. It was somewhat expected then that this native interaction would be compromised by TAR-specific mutations. This was indeed the case, and in fact two RRMs showed appreciable selectivity for TAR RNA over U1hpII. Guided by structure activity relationship precedents from the literature, we further engineered RRM mutations to shut down native binding function, while retaining TAR RNA binding activity. The final demonstration of selectivity was the recognition of TAR RNA in the presence of superstoichiometric non-specific tRNAs. This is a highly desirable feature for *in vivo* applications or *in vitro* experiments that mimic a complex biological environment.

A significant benefit to this platform is that the reagent framework can be easily modified or randomized using simple molecular biology techniques[44,49–51]. This also allows access to protein-driven methods of high-throughput screening for binding function to accelerate new reagent discovery. This is in contrast to small molecules, which require time- and labor-intensive synthesis and purification steps to make simple modifications on the carbon skeleton of a compound. Taken together, this provides a strong foundation for our alternative approach to the RNA recognition problem.

4.12 Methods

Generating Point Mutants Mutations to the U1A or ΔK50ΔM51 protein scaffolds were introduced using a common site directed mutagenesis protocol. As an example, the E19S

mutation was installed by adding ~20 ng of parent plasmid, 0.4 mM final concentration of each E19S FP (5′-CCT CAA TTC GAA GAT CAA GAA GGA TGA GCT CAA AAA GTC CC-3′) and E19S RP (5′-CCT TCT TGA TCT TCG AAT TGA GGT TGT TGA TAT AAA TAG TGT GG-3′), 0.6 mM final concentration of dNTPs and Pfu Turbo polymerase (*Stratagene*) were mixed in the appropriate reaction buffer to a final volume of 50 μL. This mixture was run through 30 cycles of polymerase chain reaction (30 seconds 90 °C, 2 minutes 56 °C, 8 minutes 72 °C) followed by 10 minutes at 72 °C. The reaction was then treated with DpnI (*New England Biolabs*) and incubated at 37 °C for a full hour. At completion, ~10 μL of the reaction was alliquoted and saved on ice while the remaining volume was run on a 1% agarose gel with a 1 k base ladder for 40-60 minutes at 120 V. Any observable amplicon of 3 k bases or larger indicated successful amplification of the plasmid. From the 10 μL aliquot, 1 μL was transformed into 5α chemically competent *E.coli* (*New England Biolabs*) according to the manufacturers instructions. After rescue for 1 hour at 37 °C, the *E.coli* were plated on LB ampicillin agar plates and incubated overnight at 37 °C. Surviving colonies were cultured in LB ampicillin for > 12 hours followed by miniprep plasmid isolation (*Omega*) and sequencing (*Genewiz*) to confirm successful molecular cloning.

Protein Expression and Purification The appropriate plasmid was transformed into BL-21 (DE-3) chemically competent *E.coli* according to the manufacturers instructions, rescued, plated, and cultured into ~10 mL of LB ampicillin. This overnight culture was inoculated into 2.5 L of LB ampicillin and incubated at 37 °C with shaking at 200 RPM (Note: at this volume,

placing a stainless steal spring in the bottom of the culture flask increases the health of the culture, leading to slightly faster growth and slightly higher protein yield). When the culture reached an OD_{600} of 0.4 to 0.8 (~3-4 hours), IPTG (*GoldBio*) was added to induce protein expression. Protein yields tend to saturate after ~ 4 hours of induction, remaining constant for at least 12 hours. The culture was centrifuged at ~10,000 RPM for 10 minutes, supernatant was discarded, and the bacterial pellet was resuspended in Hepes buffer (Hepes-NaOH 10 mM, KCl 50 mM, $MgCl_2 \cdot 6H_2O$ 1mM, NaCl 30 mM, EDTA 1 mM), 20 mL × 6. Mechanical lysis was applied by sonication, followed by centrifugation at 15,000 RPM for 10-20 minutes. Ni-NTA resin was washed with 50 mL H_2O, then 50 mL Hepes buffer. Supernatant was then collected and incubated with washed Ni-NTA resin for ~10 minutes with perturbation at 4 °C. This mixture was then centrifuged for 5 minutes, 5,000 RPM and the supernatant discarded. The resin was washed using an imidazole gradient as follows: 50 mL of 20 mM, 10 mL of 30 mM, 10 mL of 50 mM, and elution with 5 mL of 400 mM imidazole in HEPES buffer. The 400 mM imidazole elution was exhaustively dialyzed against Hepes buffer to provide pure protein for fluorescence polarization. RRMs analyzed by CD were alternatively dialyzed exhaustively against phosphate buffer (Na_2HPO_4 77.4 mM, NaH_2PO_4 22.6 mM, NaCl 30 mM, KCl 50 mM).

Fluorescence Polarization The purified RRM to be analyzed was concentrated using a centricon-4 spin column (*Millipore*) to a concentration of ~1 mM, as determined using an extinction coefficient of 7450. From this stock serial dilutions were made using a dilution factor of 1.7 to give 24 different concentrations. The appropriate fluorescein tagged nucleic acid (*Integrated DNA Technologies*, RNase free HPLC purified and shipped as a lyophilized pellet)

was thawed from an aliquoted stock of 10 μM, stored at – 80 °C. This was prepared for fluorescence polarization (FP) in a master mix containing 40 nM of indicated nucleic acid and 10% NP-40 in Hepes buffer. Prior to analysis, the mastermix was heated to 95 °C for two minutes and then plunged into ice to ensure hairpin formation. 20 μL of the protein dilutions were loaded onto a black flat-bottom 384-well plate (*Corning*) before addition of 20 μL of the RNA master mix to give 20 nM final RNA concentration. Fluorescence polarization measurements were made using a Perkin-Elmer Victor V multimode microplate reader. Data was processed using KaleidaGraph (Synergy Software) to determine RNA dissociation constants by fitting the data to single-site binding isotherm.

4.13 Proteins Used In This Work

U1A

nucleic acid sequence

ATGGCCCAGGTGCAGCTGCAGGTCGACATGGCAGTTCCCGAGACGCGTCCTAAC
CACACTATTTATATCAACAACCTCAATGAGAAGATCAAGAAGGATGAGCTCAAAA
AGTCCCTGTACGCCATCTTCTCCCAGTTTGGCCAGATCCTGGATATCCTGGTAT
CACGGAGCCTGAAGATGAGGGGCCAAGCTTTTGTCATCTTCAAGGAGGTCTCGA
GCGCCACCAACGCCCTGCGCTCCATGCAGGGTTACCCTTTCTATGACAAACCTA
TGCGTATCCAGTATGCGCGCACCGACTCAGATATCATTGCCAAGATGAAAGGCA
CCTTCGGATCGGTCGACTCTAGAGGATCCCCGGTTGCGGCCGCACATCATCACC
ATCATCACGTGGCCGCAGAACAAAAACTCATCTCAGAAGAGGATCTGAATGGGG
CCGCATAG

amino acid sequence

MAQVQLQVDMAVPETRPNHTIYINNLNEKIKKDELKKSLYAIFSQFGQILDILVSRSLK
MRGQAFVIFKEVSSATNALRSMQGYPFYDKPMRIQYARTDSDIIAKMKGTFGSVDSRGS
PVAAAHHHHHHVAAEQKLISEEDLNGAA

U1A ΔK50ΔM51

nucleic acid sequence

ATGGCCCAGGTGCAGCTGCAGGTCGACATGGCAGTTCCCGAGACGCGTCCTAAC
CACACTATTTATATCAACAACCTCAATGAGAAGATCAAGAAGGATGAGCTCAAAA
AGTCCCTGTACGCCATCTTCTCCCAGTTTGGCCAGATCCTGGATATCCTGGTAT
CACGGAGCCTGAGGGGCCAAGCTTTTGTCATCTTCAAGGAGGTCTCGAGCGCCA
CCAACGCCCTGCGCTCCATGCAGGGTTACCCTTTCTATGACAAACCTATGCGTA
TCCAGTATGCGCGCACCGACTCAGATATCATTGCCAAGATGAAAGGCACCTTCG
GATCGGTCGACTCTAGAGGATCCCCGGTTGCGGCCGCACATCATCACCATCATC
ACGTGGCCGCAGAACAAAAACTCATCTCAGAAGAGGATCTGAATGGGGCCGCAT
AG

amino acid sequence

MAQVQLQVDMAVPETRPNHTIYINNLNEKIKKDELKKSLYAIFSQFGQILDILVSRSLRG
QAFVIFKEVSSATNALRSMQGYPFYDKPMRIQYARTDSDIIAKMKGTFGSVDSRGSPVA
AAHHHHHHVAAEQKLISEEDLNGAA

4.14 Primers Used in This Work

Primer Name	Sequence
FP U1A N15A	5′-CCTAACCACACTATTTATATCGCCAACCTCAATGAGAAGATCAAG-3′
RP U1A N15A	5′-CTTGATCTTCTCATTGAGGTTGGCGATATAAATAGTGTGGTTAGG-3′
FP U1A N16A	5′-CCTAACCACACTATTTATATCAACGCCCTCAATGAGAAGATCAAGAAGG-3′
RP U1A N16A	5′-CCTTCTTGATCTTCTCATTGAGGGCGTTGATATAAATAGTGTGGTTAGG-3′
FP U1A E19A	5′-CCACACTATTTATATCAACAACCTCAATGCGAAGATCAAGAAGGATGAGCTC-3′
RP U1A E19A	5′-GAGCTCATCCTTCTTGATCTTCGCATTGAGGTTGTTGATATAAATAGTGTGG-3′
FP U1A S46A	5′-GATATCCTGGTAGCACGGAGCCTGAAG-3′
RP U1A S46A	5′-CTCCGTGCTACCAGGATATCCAGGATC-3′
FP U1A S48A	5′-ACGGGCCCTGAAGATGAGGGGC-3′
RP U1A S48A	5′-CATCTTCAGGGCCCGTGATACCAGGATATCCAGG-3′
FPU1A L49A	5′-GAGCGCGAAGATGAGGGGCCAAG-3′
RP U1A L49A	5′-CCTCATCTTCGCGCTCCGTGATACCAGG-3′
FP U1A K50A	5′-GCCTGGCGATGAGGGGCC-3′
RP U1A K50A	5′-CCTCATCGCCAGGCTCCGTGATACC-3′
FP U1A M51A	5′-AGCCTGAAGGCGAGGGGCC-3′
RP U1A M51A	5′-CCTCGCCTTCAGGCTCCGTGATACCAG-3′
FP ΔKΔM S46A	5′-GATATCCTGGTAGCACGGAGCCTGAGG-3′
RP ΔKΔM S46A	5′-CTCCGTGCTACCAGGATATCCAGGATC-3′
FP ΔKΔM S48A	5′-GGTATCACGGGCCCTGAGGGGCC-3′
RP ΔKΔM S48A	5′-CTCAGGGCCCGTGATACCAGGATATCCAGGC-3′
FP ΔKΔM L49A	5′-CGGAGCGCGAGGGGC-3′
RP ΔKΔM L49A	5′-CCTCGCGCTCCGTGATACCAGG-3′
FP U1A E19S	5′-CCTCAATTCGAAGATCAAGAAGGATGAGCTCAAAAAGTCCC-3′
RP U1A E19S	5′-CCTTCTTGATCTTCGAATTGAGGTTGTTGATATAAATAGTGTGG-3′
FP U1A E19F	5′-CCTCAATTTCAAGATCAAGAAGGATGAGCTCAAAAAGTCCC-3′
RP U1A E19F	5′-CCTTCTTGATCTTGAAATTGAGGTTGTTGATATAAATAGTGTGG-3′
FP U1A E19K	5′-CCTCAATAAGAAGATCAAGAAGGATGAGCTCAAAAAGTCCC-3′
RP U1A E19K	5′-CCTTCTTGATCTTCTTATTGAGGTTGTTGATATAAATAGTGTGG-3′
FP U1A E19Q	5′-CCTCAATCAGAAGATCAAGAAGGATGAGCTCAAAAAGTCCC-3′
RP U1A E19Q	5′-CCTTCTTGATCTTCTGATTGAGGTTGTTGATATAAATAGTGTGG-3′
FP U1A L49E	5′-GAGCGAGAAGATGAGGGGCCAAG-3′
RP U1A L49E	5′-CCTCATCTTCTCGCTCCGTGATACCAGG-3′
FP U1A L49F	5′-GAGCTTCAAGATGAGGGGCCAAG-3′
RP U1A L49F	5′-CCTCATCTTGAAGCTCCGTGATACCAGG-3′
FP U1A L49K	5′-GAGCAAGAAGATGAGGGGCCAAG-3′
RP U1A L49K	5′-CCTCATCTTCTTGCTCCGTGATACCAGG-3′
FP U1A L49N	5′-GAGCAACAAGATGAGGGGCCAAG-3′
RP U1A L49N	5′-CCTCATCTTGTTGCTCCGTGATACCAGG-3′
FP U1A L49S	5′-GAGCAGCAAGATGAGGGGCCAAG-3′
RP U1A L49S	5′-CCTCATCTTGCTGCTCCGTGATACCAGG-3′
FP U1A M51E	5′-AGCCTGAAGGAGAGGGGCC-3′
RP U1A M51E	5′-CCTCTCCTTCAGGCTCCGTGATACCAG-3′
FP U1A M51F	5′-AGCCTGAAGTTCAGGGGCC-3′
RP U1A M51F	5′-CCTGAACTTCAGGCTCCGTGATACCAG-3′
FP U1A M51K	5′-AGCCTGAAGAAGAGGGGCC-3′
RP U1A M51K	5′-CCTCTTCTTCAGGCTCCGTGATACCAG-3′
FP U1A M51N	5′-AGCCTGAAGAACAGGGGCC-3′
RP U1A M51N	5′-CCTGTTCTTCAGGCTCCGTGATACCAG-3′
FP U1A M51S	5′-AGCCTGAAGAGCAGGGGCC-3′
RP U1A M51S	5′-CCTGCTCTTCAGGCTCCGTGATACCAG-3′
FP U1A Y13Q	5′-CGCGTCCTAACCACACTATTGGTATCAACAACCTCAATGAGAAGATC-3′
RP U1A Y13Q	5′-GGTTGTTGATCTGAATAGTGTGGTTAGGACGCGTCTCGG-3′
FP U1A F56A	5′-CGGAGCCTGAGGGGCCAAGCCGCGGTCATCTTCAAGGAGG-3′
RP U1A F56A	5′-CCTCCTTGAAGATGACCGCGGCTTGGCCCCTCAGGCTCCG-3′
FP U1A N15V	5′-CCACACTATTTATATCGTCAACCTCAATGAGAAGATCAAGAAGGATGAGCTCAAAAAGTC-3′
RP U1A N15V	5′-GAGGTTGACGATATAAATAGTGTGGTTAGGACGCGTCTCGGG-3′

FP U1A L49A M51A	5′-GAGCGCGAAGGCGAGGGGCCAAGCTTTTGTC-3′
RP U1A L49A M51A	5′-CCCTCGCCTTCGCGCTCCGTGATACCAGG-3′
FP U1A S48A L49A	5′-ACGGGCCGCGAAGATGAGGGGCCAAGC-3′
RP U1A S48A L49A	5′-CATCTTCGCGGCCCGTGATACCAGGATATCCAGG-3′
FP U1A S48A M51A	5′-ACGGGCCCTGAAGGCGAGGGGCCAAGCTTTTGTC-3′
RP U1A S48A M51A	5′-CCCTCGCCTTCAGGGCCCGTGATACCAGG-3′
FP U1A S48A L49A M51A	5′-ACGGGCCGCGAAGGCGAGGGGCCAAGCTTTTGTC-3′
RP U1A S48A L49A M51A	5′-CCCTCGCCTTCGCGGCCCGTGATACCAGGATATCCAGG-3′
FP U1A N15A N16A	5′-CTATTTATATCGCCGCCCTCAATGAGAAGATCAAGAAGGATGAGCTCAAAAAGTCC-3′
RP U1A N15A N16A	5′-GAGGGCGGCGATATAAATAGTGTGGTTAGGACGCGTCTCGGG-3′
FP U1A N15A E19A	5′-GCCAACCTCAATGCGAAGATCAAGAAGGATGAGCTCAAAAAGTCCC-3′
RP U1A N15A E19A	5′-CGCATTGAGGTTGGCGATATAAATAGTGTGGTTAGGACGCGTCTCGG-3′
FP U1A N16A E19A	5′-ACGCCCTCAATGCGAAGATCAAGAAGGATGAGCTCAAAAAGTCCC-3′
RP U1A N16A E19A	5′-CGCATTGAGGGCGTTGATATAAATAGTGTGGTTAGGACGCGTCTCG-3′
FP U1A Y13Q E19S	5′-CCACACTATTCAGATCAACAACCTCAATTCGAAGATCAAGAAGGATGAGC-3′
FP U1A Y13Q E19A	5′-CCACACTATTCAGATCAACAACCTCAATGCGAAGATCAAGAAGGATGAGC-3′
FP U1A Y13Q E19F	5′-CCACACTATTCAGATCAACAACCTCAATTTCAAGATCAAGAAGGATGAGC-3′
FP U1A N15V E19S	5′-CCACACTATTTATATCGTCAACCTCAATTCGAAGATCAAGAAGGATGAGCTCAAAAAGTC-3′
FP U1A N15V E19A	5′-CCACACTATTTATATCGTCAACCTCAATGCGAAGATCAAGAAGGATGAGCTCAAAAAGTC-3′
FP U1A N15V E19F	5′-CCACACTATTTATATCGTCAACCTCAATTTCAAGATCAAGAAGGATGAGCTCAAAAAGTC-3′

4.15 Nucleic Acids Used In This Work

Nucleic Acid	Sequence
TAR RNA	5′-GGCAGAUCUGAGCCUGGGAGCUCUCUGCC-3′
U1hpII RNA	5′-AGCTTATCCATTGCACCGGATAAGCT-3′
TAR DNA	5′-GGCAGATCTGAGCCTGGGAGCTCTCTGCC-3′

4.16 Supplemental Data

Figure S4.1 Fluorescence polarization isotherms for binding data presented in **Table 4.1**. Binding affinities for U1A or ΔK50ΔM51, or indicated alanine mutants thereof, in complex with TAR RNA.

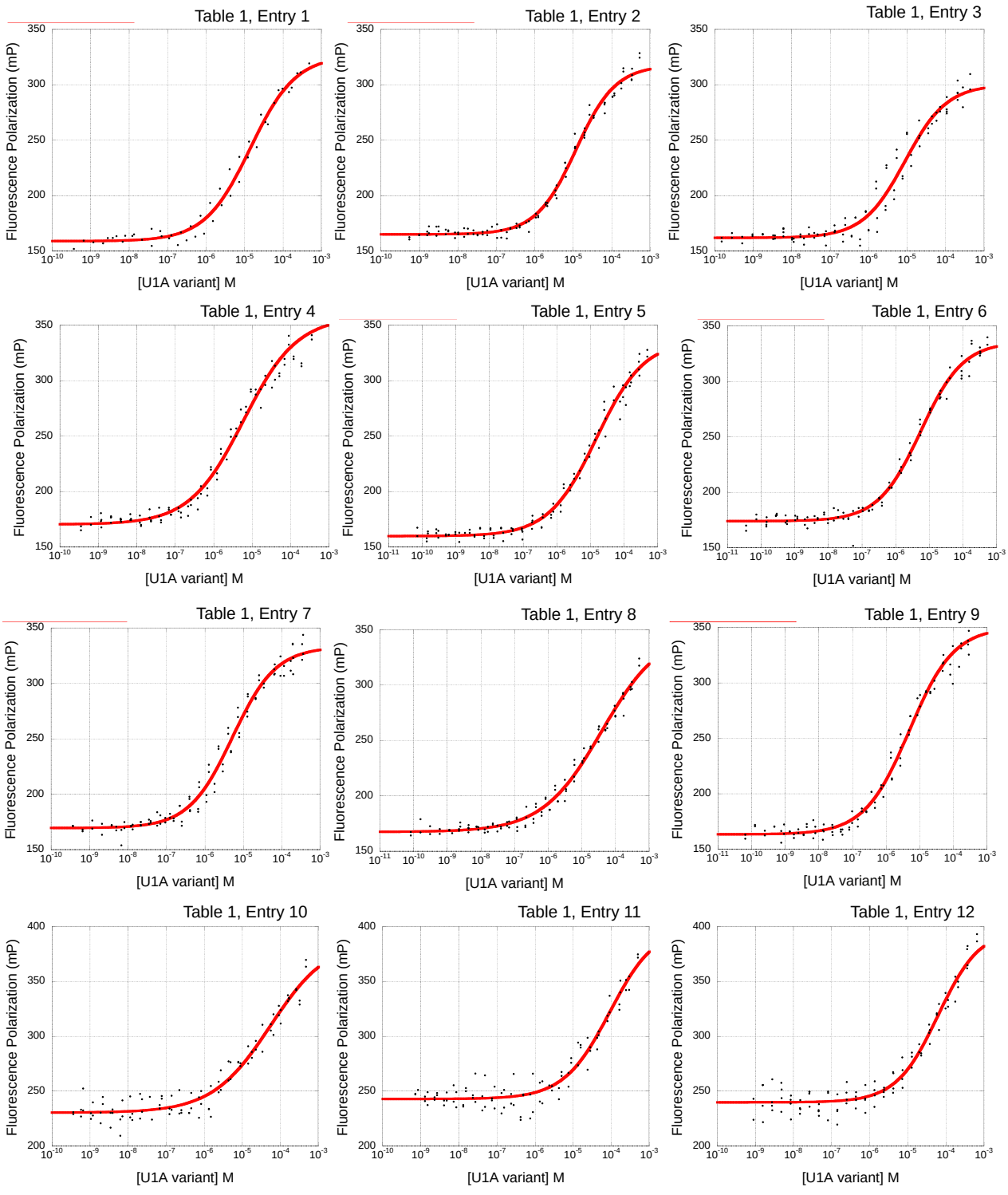


Table 1, Entry 1
Table 1, Entry 2
Table 1, Entry 3
Table 1, Entry 4
Table 1, Entry 5
Table 1, Entry 6
Table 1, Entry 7
Table 1, Entry 8
Table 1, Entry 9
Table 1, Entry 10
Table 1, Entry 11
Table 1, Entry 12
Fluorescence Polarization (mP)
[U1A variant] M

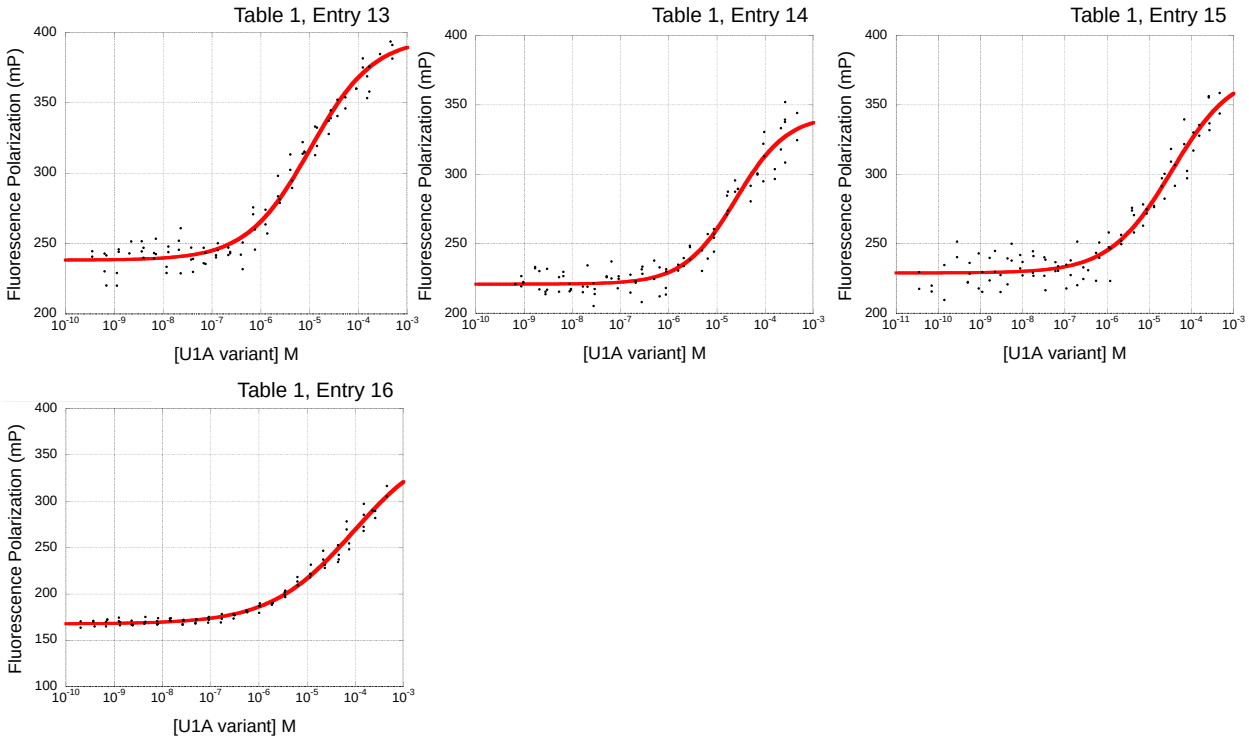
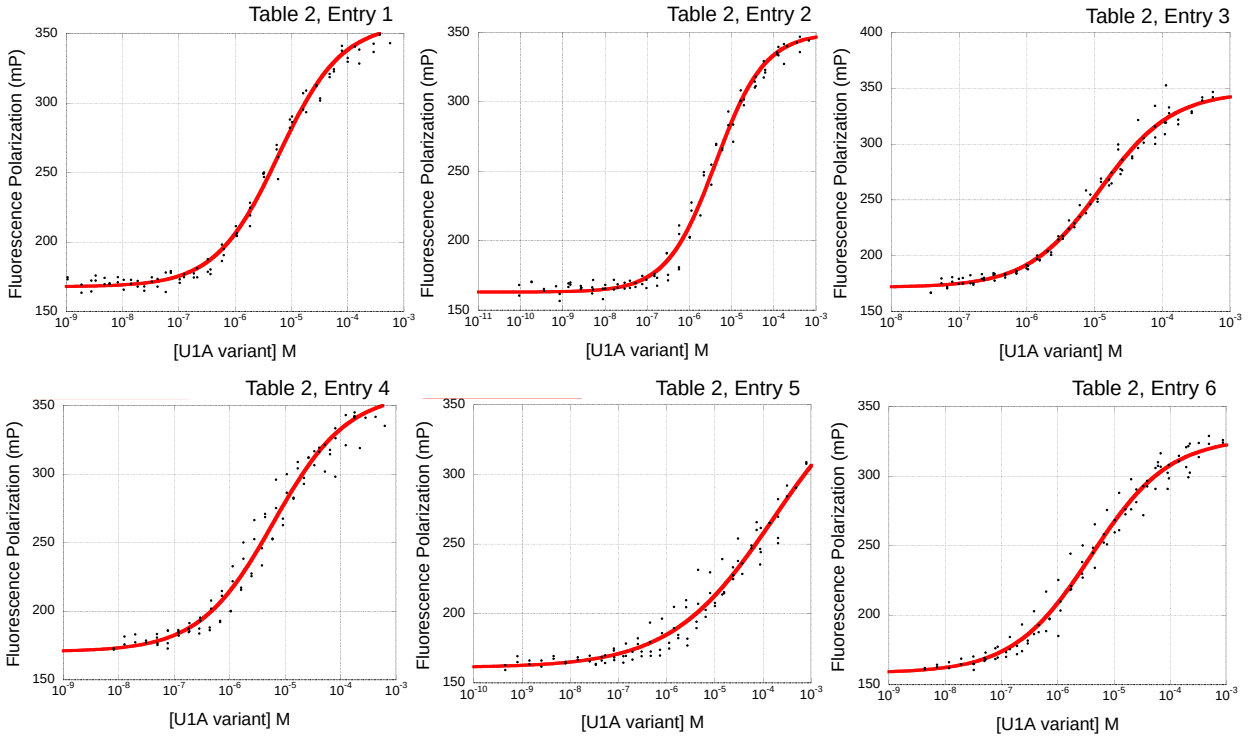


Figure S4.2 Fluorescence polarization isotherms for binding data presented in **Table 4.2**. Binding affinities for U1A or ΔK50ΔM51, or indicated point mutants thereof, in complex with TAR RNA.



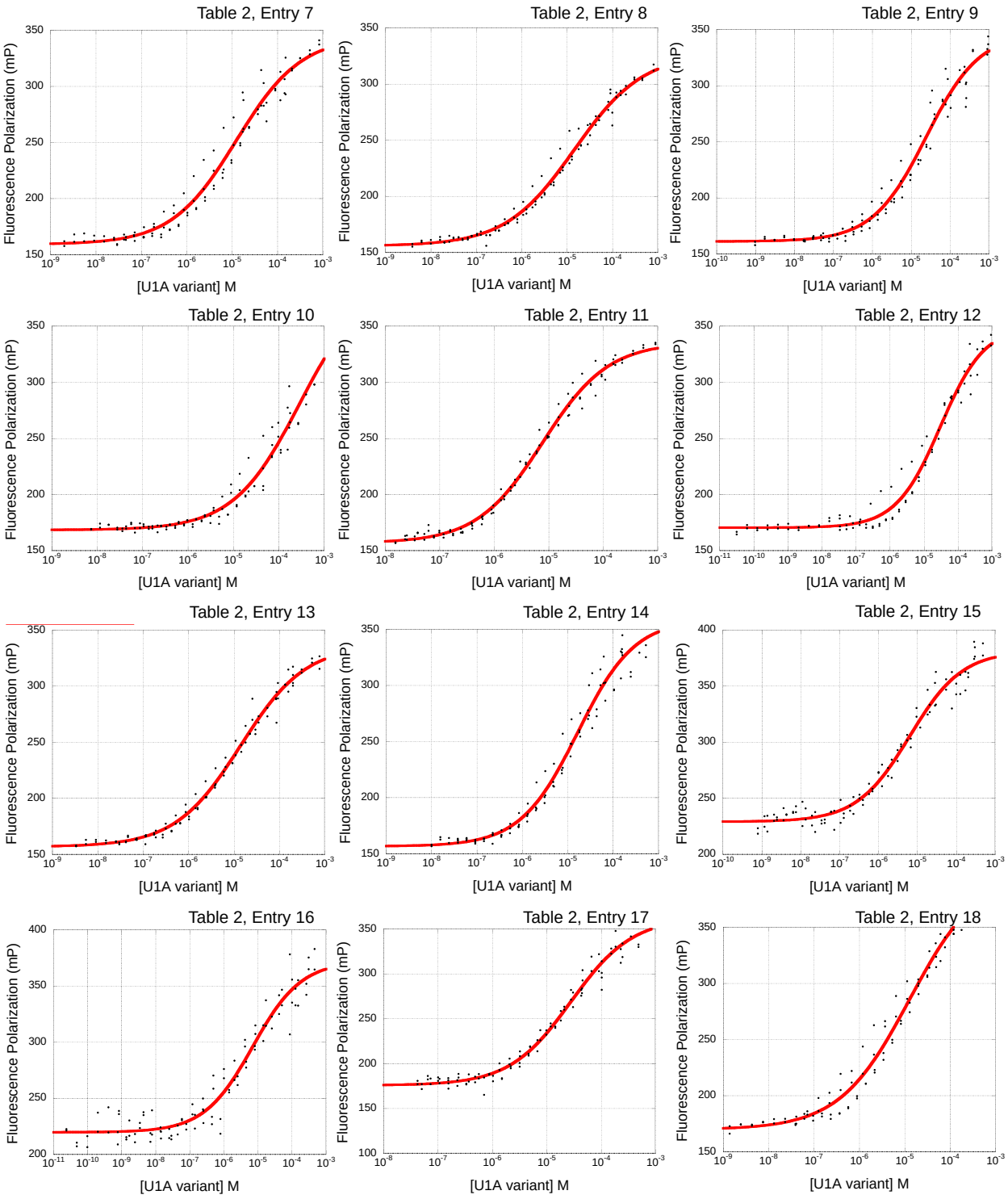


Figure S4.3 Fluorescence polarization isotherms for binding data presented in **Table 4.3**. Binding affinities for U1A or ΔK50ΔM51 mutants in complex with TAR DNA or U1hpII RNA.

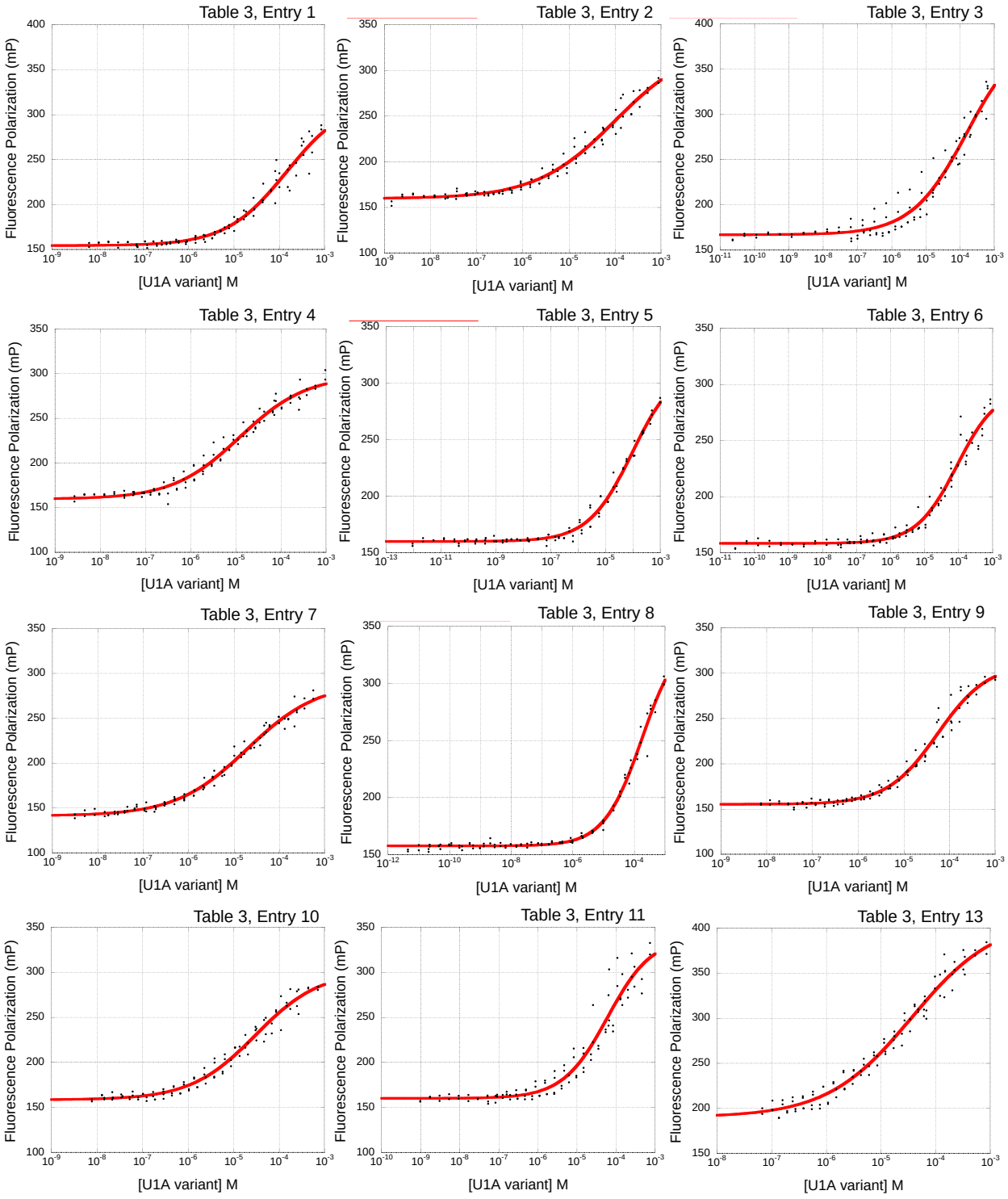


Table 3, Entry 1
Table 3, Entry 2
Table 3, Entry 3
Table 3, Entry 4
Table 3, Entry 5
Table 3, Entry 6
Table 3, Entry 7
Table 3, Entry 8
Table 3, Entry 9
Table 3, Entry 10
Table 3, Entry 11
Table 3, Entry 13
Fluorescence Polarization (mP)
[U1A variant] M

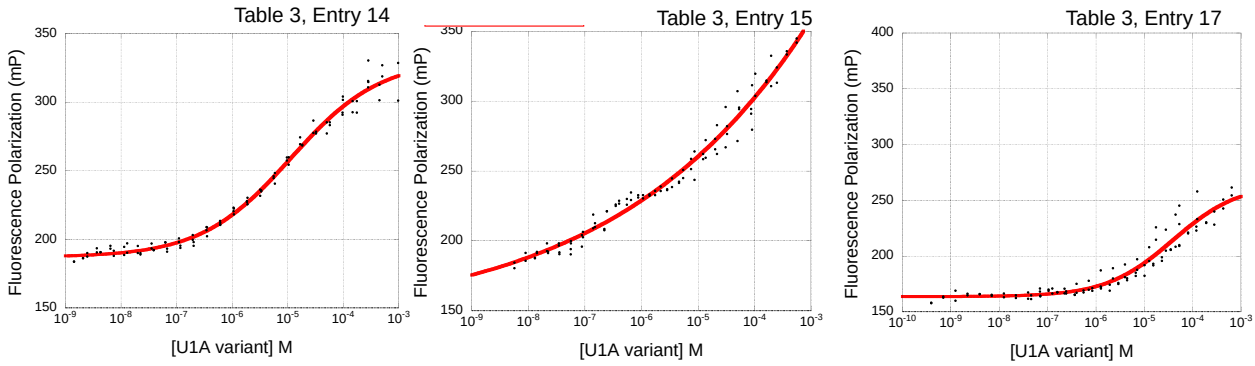
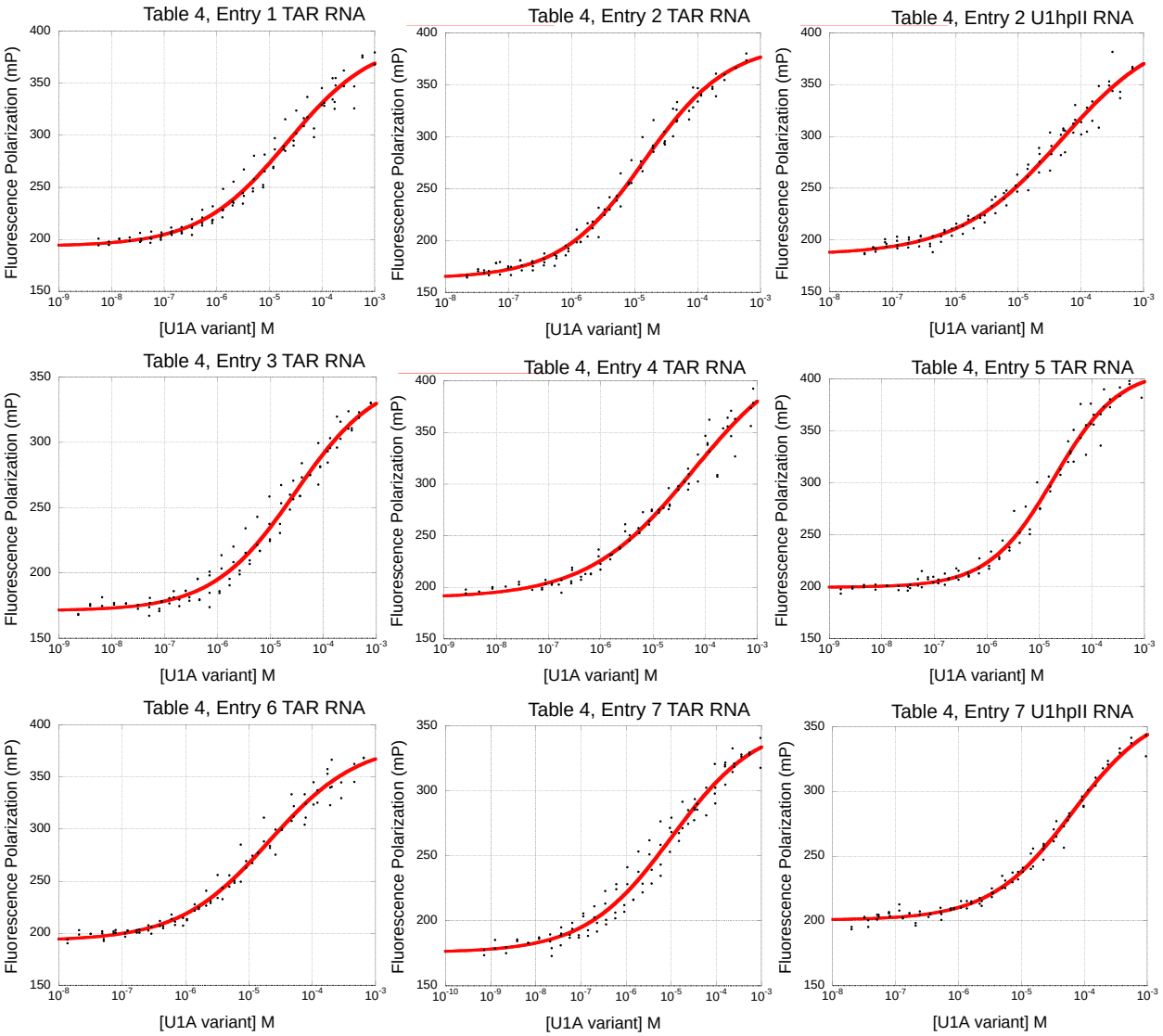


Figure S4.4 Fluorescence polarization isotherms for binding data presented in **Table 4.4**. Binding affinities for U1A double mutants in complex with TAR RNA or U1hpII RNA.



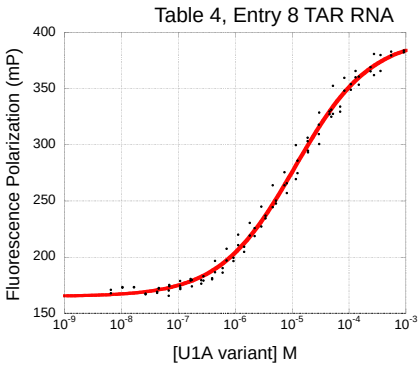


Table 4, Entry 8 TAR RNA
Fluorescence Polarization (mP)
[U1A variant] M

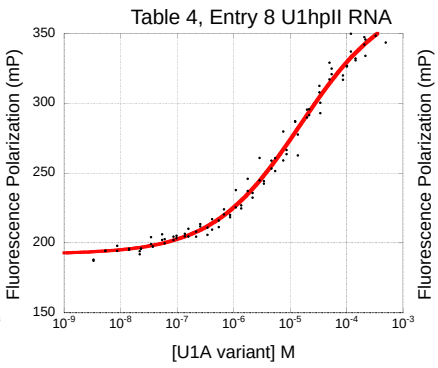


Table 4, Entry 8 U1hpII RNA
Fluorescence Polarization (mP)
[U1A variant] M

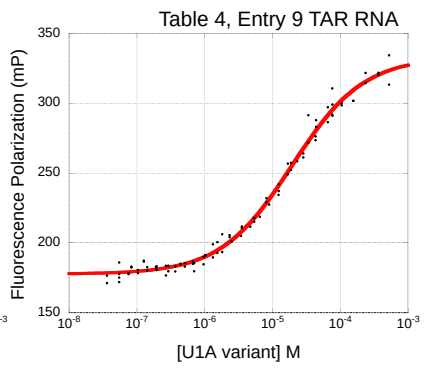


Table 4, Entry 9 TAR RNA
Fluorescence Polarization (mP)
[U1A variant] M

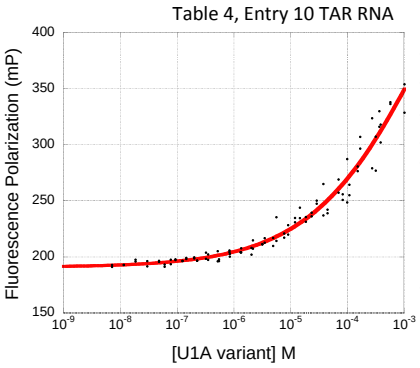


Table 4, Entry 10 TAR RNA
Fluorescence Polarization (mP)
[U1A variant] M

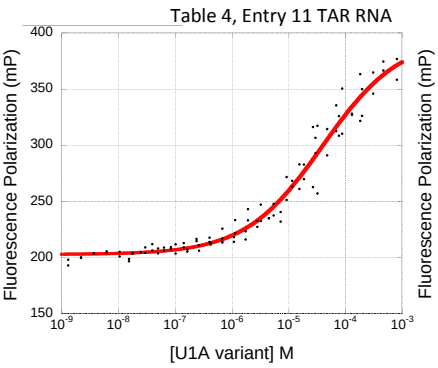


Table 4, Entry 11 TAR RNA
Fluorescence Polarization (mP)
[U1A variant] M

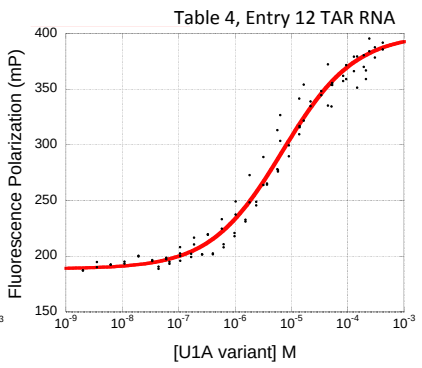


Table 4, Entry 12 TAR RNA
Fluorescence Polarization (mP)
[U1A variant] M

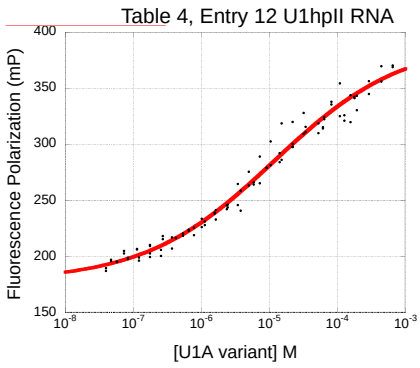


Table 4, Entry 12 U1hpII RNA
Fluorescence Polarization (mP)
[U1A variant] M

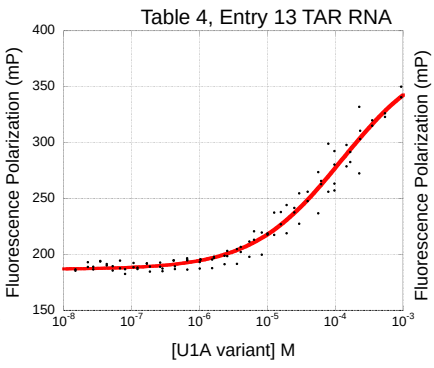


Table 4, Entry 13 TAR RNA
Fluorescence Polarization (mP)
[U1A variant] M

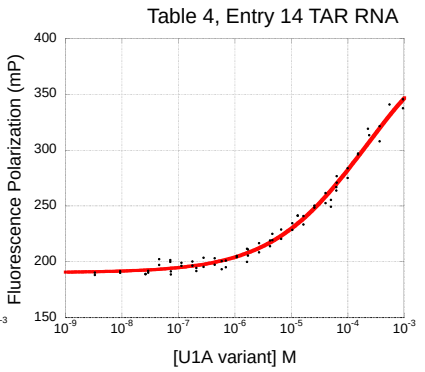


Table 4, Entry 14 TAR RNA
Fluorescence Polarization (mP)
[U1A variant] M

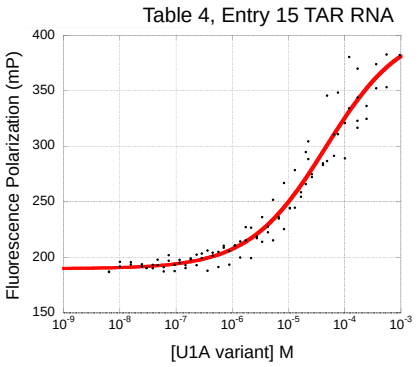


Table 4, Entry 15 TAR RNA
Fluorescence Polarization (mP)
[U1A variant] M

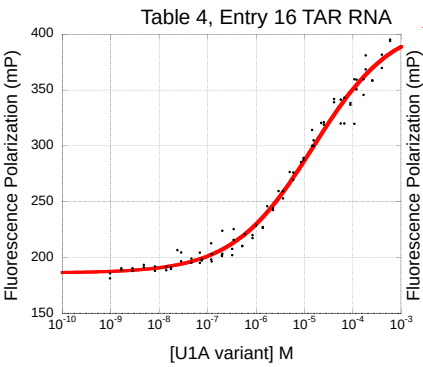


Table 4, Entry 16 TAR RNA
Fluorescence Polarization (mP)
[U1A variant] M

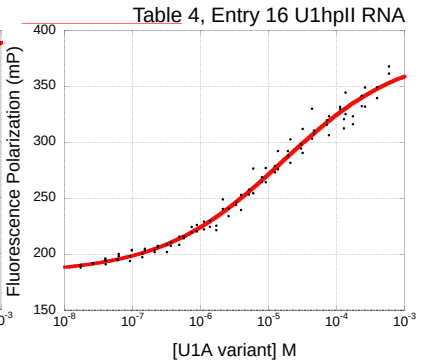


Table 4, Entry 16 U1hpII RNA
Fluorescence Polarization (mP)
[U1A variant] M

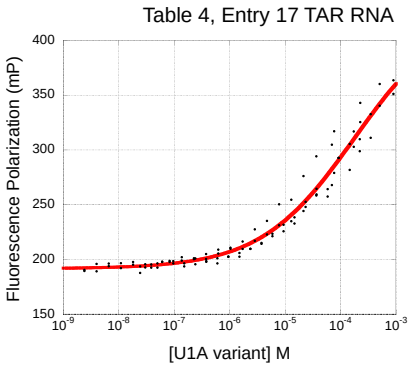


Figure S4.5 Fluorescence polarization isotherms for binding data presented in **Table 4.5**. binding affinities for U1A double mutants in comples with TAR RNA in the presence of 10 molar excess tRNAs.

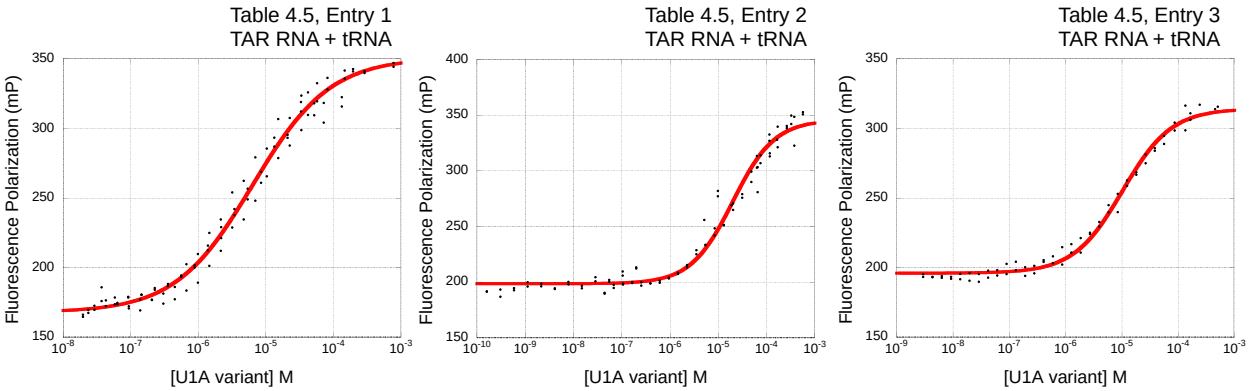
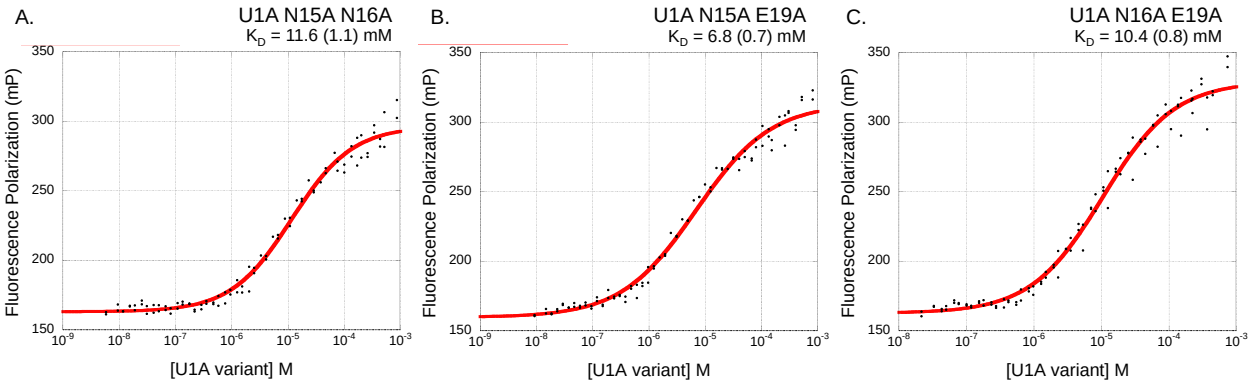


Figure S4.6 Fluorescence polarization isotherms for RRMs with combined beneficial mutations in complex with TAR RNA.



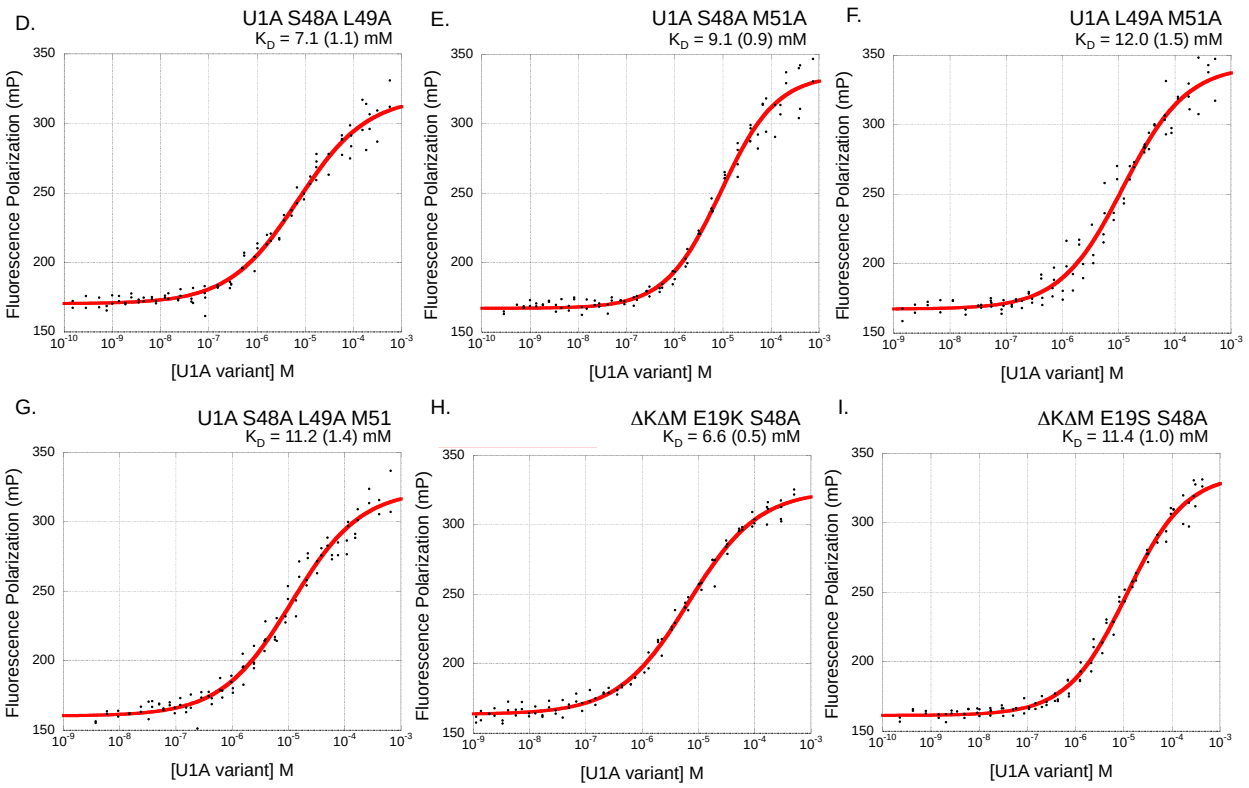
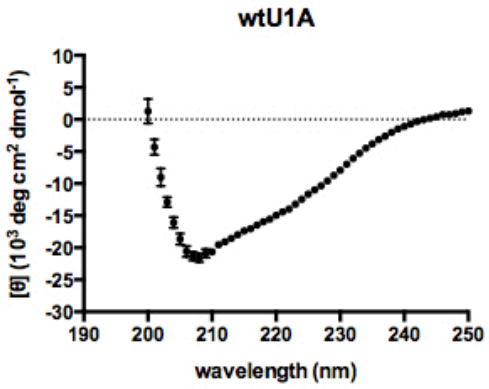
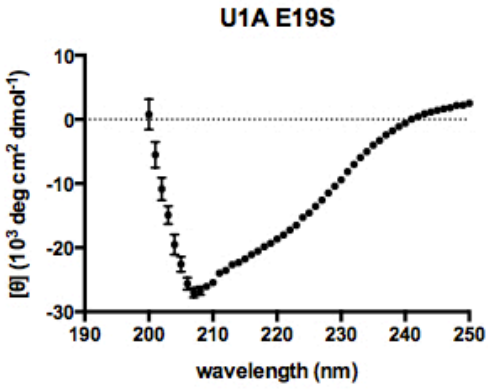
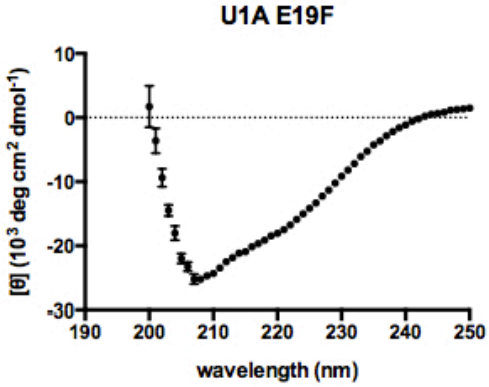
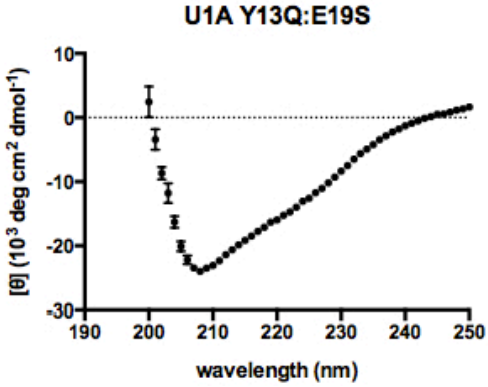


Figure S4.7 Circular dichroism data for U1A, U1A_E19S, U1A_E19F, and U1A_Y13Q:E19S.


wtU1A
[θ] (10³ deg cm² dmol⁻¹)
wavelength (nm)


U1A E19S
[θ] (10³ deg cm² dmol⁻¹)
wavelength (nm)


U1A E19F
[θ] (10³ deg cm² dmol⁻¹)
wavelength (nm)


U1A Y13Q:E19S
[θ] (10³ deg cm² dmol⁻¹)
wavelength (nm)

REFERENCES

(1) Cooper, T.; Wan, L.; Dreyfuss, G. *Cell* **2009**, *136*, 777.

(2) Osborne, R. J.; Thornton, C. *Hum. Mol. Genet.* **2006**, *15 Spec No*, R162.

(3) Esteller, M. *Nat. Rev. Genet.* **2011**, *12*, 861.

(4) Costa, F. F. *BioEssays* **2010**, *32*, 599.

(5) Ling, H.; Fabbri, M.; Calin, G. A. *Nat. Rev. Drug Discov.* **2013**.

(6) Blakeley, B. D.; Shattuck, J.; Coates, M. B.; Tran, E.; Laird-Offringa, I. A.; McNaughton, B. R. *Biochemistry* **2013**, *52*, 4745.

(7) Oubridge, C.; Ito, N.; Evans, P. R.; Teo, C. H.; Nagai, K. *Nature* **1994**, *372*, 432.

(8) Kranz, J. K.; Hall, K. B. *J. Mol. Biol.* **1999**, *285*, 215.

(9) Law, M. J.; Rice, A. J.; Lin, P.; Laird-Offringa, I. a. *RNA* **2006**, *12*, 1168.

(10) Katsamba, P. S.; Bayramyan, M.; Haworth, I. S.; Myszka, D. G.; Laird-Offringa, I. a. *J. Biol. Chem.* **2002**, *277*, 33267.

(11) Kranz, J. K.; Hall, K. B. *J. Mol. Biol.* **1998**, *275*, 465.

(12) Allain, F. H.; Howe, P. W.; Neuhaus, D.; Varani, G. *EMBO J.* **1997**, *16*, 5764.

(13) Yan Fan. *Thesis Dissertation* University of Illinois at Urbana-Champaign **2009**.

(14) Nolan, S. J.; Shiels, J. C.; Tuite, J. B.; Cecere, K. L.; Baranger, A. M. *J Am Chem Soc* **1999**, 8951.

(15) Scherly, D.; Boelens, W.; Dathan, N. A.; van Venrooij, W. J.; Mattaj, I. W. *Nature* **1990**, *345*, 502.

(16) Bentley, R. C.; Keene, J. D. *Mol. Cell. Biol.* **1991**.

(17) Law, M. J.; Lee, D. S.; Lee, C. S.; Anglim, P. P.; Haworth, I. S.; Laird-Offringa, I. A. *Nucleic Acids Res.* **2013**, *41*, 7092.

(18) Laird-Offringa, I. A.; Belasco, J. G. *Proc. Natl. Acad. Sci.* **1993**, *92*, 11859.

(19) Law, M. J.; Chambers, E. J.; Katsamba, P. S.; Haworth, I. S.; Laird-Offringa, I. a. *Nucleic Acids Res.* **2005**, *33*, 2917.

(20) Rana, T. M.; Jeang, K. *Archives of Biochem Biophys Minirev* **1999**, *365*, 175.

(21) Maitra, R. K.; McMillion, N. A. J.; Desai, S.; McSwiggen, J.; Hovanessian, A. G.; Ganes, S.; Willioams, B. R. G.; Silverman, Robert, H. *Virol* **1994**, *204*, 823.

(22) Berkhout, B.; Gatignol, a; Rabson, a B.; Jeang, K. T. *Cell* **1990**, *62*, 757.

(23) Dingwall, C.; Ernberg, I.; Gait, M. J.; Green, S. M.; Heaphy, S.; Karn, J.; Lowe, a D.; Singh, M.; Skinner, M. a. *EMBO J.* **1990**, *9*, 4145.

(24) Athanassiou, Z.; Dias, R. L. A.; Moehle, K.; Dobson, N.; Varani, G.; Robinson, J. A. *J. Am. Chem. Soc.* **2004**, *126*, 6906.

(25) Leeper, T. C.; Athanassiou, Z.; Dias, R. L. A.; Robinson, J. A.; Varani, G. *Biochemistry* **2005**, *44*, 12362.

(26) Chittapragada, M.; Roberts, S.; Ham, Y. W. *Perspect. Medicin. Chem.* **2009**, *3*, 21.

(27) Bardaro, M. F.; Shajani, Z.; Patora-Komisarska, K.; Robinson, J. A.; Varani, G. *Nucleic Acids Res.* **2009**, *37*, 1529.

(28) Davidson, A.; Patora-Komisarska, K.; Robinson, J. A.; Varani, G. *Nucleic Acids Res.* **2011**, *39*, 248.

(29) Churcher, M. J.; Lamont, C.; Hamy, F.; Dingwall, C.; Green, S. M.; Lowe, A. D.; Butler, P. J. G.; Gait, M. J.; Karn, J. *J. Mol. Biol.* **1993**, *230*, 90.

(30) Jeang, K.; Chun, R.; Lin, N. H.; Gatignol, A.; Glabe, C. G.; Fan, H. *J Virol* **1993**, *67*, 6224.

(31) Hamy, F.; Felder, E. R.; Heizmann, G.; Lazdins, J.; Aboul-ela, F.; Varani, G.; Karn, J.; Klimkait, T. *Proc. Natl. Acad. Sci. U. S. A.* **1997**, *94*, 3548.

(32) Allain, F. H.; Howe, P. W.; Neuhaus, D.; Varani, G. *EMBO J.* **1997**, *16*, 5764.

(33) Jessen, T. H.; Oubridge, C.; Teo, C. H.; Pritchard, C.; Nagai, K. *EMBO J.* **1991**, *10*, 3447.

(34) Qin, F.; Chen, Y.; Wu, M.; Li, Y.; Zhang, J.; Chen, H. F. *RNA* **2010**, *16*, 1053.

(35) Showalter, S. a; Hall, K. B. *J. Mol. Biol.* **2004**, *335*, 465.

(36) Benitex, Y.; Baranger, A. M. *BMC Biochem.* **2007**, *8*, 22.

(37) Walter, F.; Vicens, Q.; Westhof, E. *Curr. Opin. Chem. Biol.* **1999**, *3*, 694.

(38) Hainrichson, M.; Nudelman, I.; Baasov, T. *Org. Biomol. Chem.* **2008**, *6*, 227.

(39) Paul, D. J.; Seedhouse, S. J.; Disney, M. D. *Nucleic Acids Res.* **2009**, *37*, 5894.

(40) Gallego, J.; Varani, G. *Acc Chem Res* **2001**, *34*, 836.

(41) Hermann, T. *Biopolymers* **2003**, *70*, 4.

(42) Reetz, M. T.; Carballeira, J. D. *Nat. Protoc.* **2007**, *2*, 891.

(43) Farinas, E. T.; Bulter, T.; Arnold, F. H. *Curr. Opin. Biotechnol.* **2001**, *12*, 545.

(44) Romero, P. A.; Arnold, F. H. *Nat. Rev. Mol. Cell Biol.* **2009**, *10*, 866.

(45) Liang, F.-S.; Greenberg, W. A.; Hammond, J. A.; Hoffmann, J.; Head, S. R.; Wong, C.-H. *Proc. Natl. Acad. Sci.* **2006**, *103*, 12311.

(46) Luedtke, N. W.; Liu, Q.; Tor, Y. *Biochemistry* **2003**, *42*, 11391.

(47) Vester, B.; Wengel, J. *Biochemistry* **2004**, *43*, 13233.

(48) Thomas, J. R.; Liu, X.; Hergenrother, P. J. *Biochemistry* **2006**, *45*, 10928.

(49) Brustad, E. M.; Arnold, F. H. *Curr. Opin. Chem. Biol.* **2011**, *15*, 201.

(50) Boder, E. T.; Wittrup, K. D. *Nat. Biotechnol.* **1997**, *15*, 553.

(51) Bazan, J.; Całkosiński, I.; Gamian, A. *Hum. Vaccin. Immunother.* **2012**, *8*, 1817.

CHAPTER FIVE
YEAST DISPLAY MATURATION OF SYNTHETIC RNA RECOGNITION MOTIFS THAT POTENTLY AND SELECTIVELY BIND TAR RNA

5.1 Introduction

In the quest to develop a strategy for sequence-selective recognition of RNA hairpins, we have championed the use of existing RNA-binding proteins (RNA Recognition Motifs, RRMs) as scaffolds[1,2]. The RRM protein U1A embodies a scaffold that incorporates the folded three-dimensional structure of its RNA target into the recognition of its cognate RNA binding partner, U1hpII[3–6] **Figure 5.1A**. Previously, we have shown that directed mutations to the surface of the U1A scaffold could change its RNA binding specificity. Presented here are efforts to expand on those findings by exploring a larger area of U1A scaffold sequence space. Using high-throughput screening, we are able to precisely tune the affinity and specificity of this scaffold. This work is informed by previous biophysical and mutational studies by other researchers on the native interface.

In the native interaction with U1hpII hairpin RNA, multiple regions, or sub-domains, of U1A undergo conformational changes or are pre-positioned to engage in productive contacts with U1hpII. The β2-β3 loop of U1A is disordered, but in complex with U1hpII adopts a defined loop feature that fits into the hairpin loop of U1hpII[7–10] **Figure 5.1A**. Modifications to this loop have been shown to instigate new RNA hairpin loop binding specificities. New binding activity of β2-β3 loop mutants include increased affinity for U1hpII[11], improving the protein's

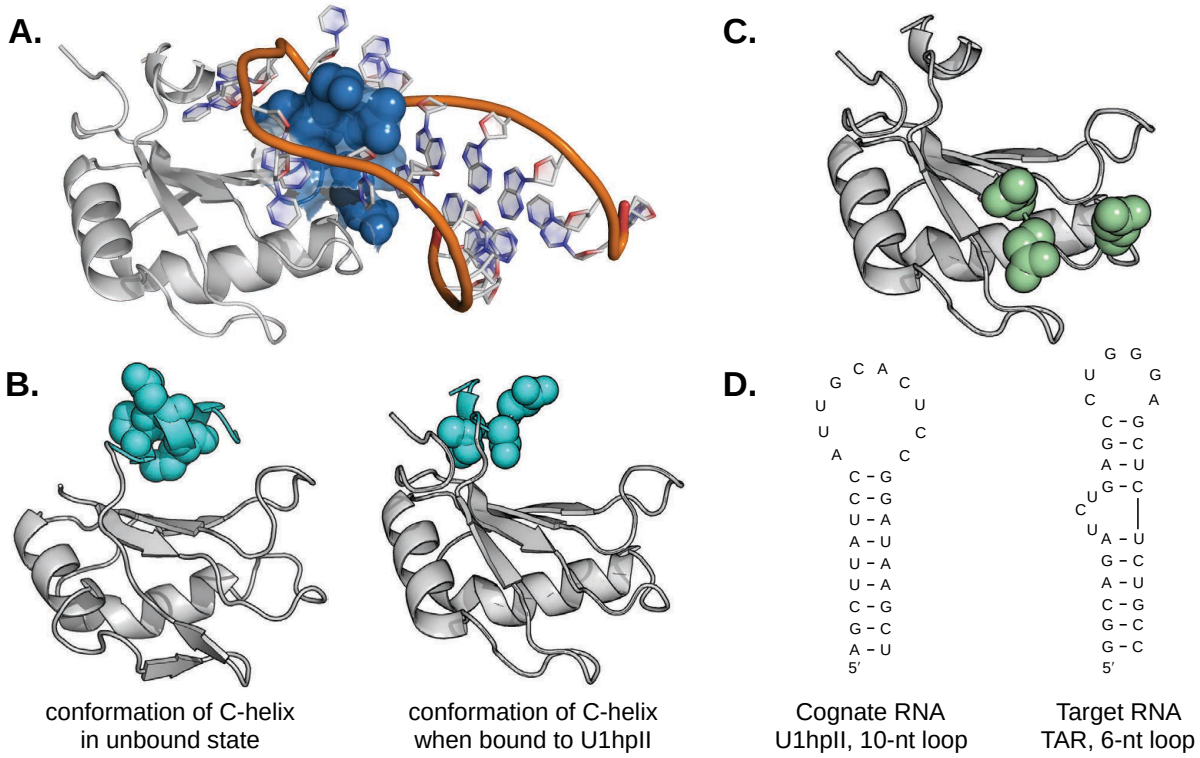


Figure 5.1 A The β2-β3 loop of U1A (dark blue spheres) fits into the hairpin loop of U1hpII in the native complex. This region contributes significantly to the shape complementarity to the RNA binding partner. **B** The C-terminal helix of U1A (cyan) rearranges to expose an RNA binding trough to accommodate the U1hpII loop. **C** Residues on the β1-α1 loop (green) are specifically positioned on the U1A scaffold to directly engage loop bases on U1hpII. **D** Nucleotide sequence and Watson-Crick interactions of the cognate RNA, U1hpII, and the new target RNA, TAR RNA. The TAR RNA does not share loop size or sequence homology with U1hpII. (PDB codes: U1A bound to U1hpII, 1URN and U1A unbound, no RNA 1FHT)

tolerance of smaller hairpin loops[1,10] **Chapter 3** and **Figure 4.2A**, and redirecting affinity to another hairpin RNA with good size and sequence homology to U1hpII[8] **Figure 4.2B**. These examples in conjunction with our own analysis have led us to the hypothesis that the β2-β3 loop on U1A primarily functions to improve the shape complementarity to its RNA binding partner. Collectively, these observations strongly suggest that this sub-domain is critical for determining RNA binding specificity within the U1A scaffold.

Other researchers have shown a second conformational change, repositioning of the C-helix sub-domain, in U1A upon binding U1hpII[12–1415] **Figure 5.1B**. This shift exposes a

hydrophobic trough on the face of the protein, in which U1hpII hairpin loop bases are buried against aromatic residues of the protein **Figure 3.2A**. Additionally, Laird-Offringa and coworkers have demonstrated a significant contribution from residues Ser91, Asp92, and Ile94 to the binding interaction[14] **Figure 5.1B**. Together, the β2-β3 loop and the C-helix appear to "clamp" the loop of the RNA hairpin, holding it in the hydrophobic trough. A third region, the β1-α1 loop, does not undergo an appreciable conformational change but instead relies on the positioning of the bound RNA on the scaffold in order to engage hairpin loop nucleotides[16–18] **Figure 5.1C**. Residues on this loop present hydrogen bonding and accepting functional groups for interaction with the bases on these nucleotides **Figure 3.2B**. Collectively, each of these regions sculpts a topology that conforms to the hairpin shape of U1hpII, enhancing precise contacts in the interaction.

Our previous work strongly suggests that U1A is a privileged protein scaffold for RNA hairpin recognition[2]. This affords compelling opportunities as, unlike privileged natural product or small molecule scaffolds, the chemical functionality of proteins is easily modified[19,20]. A parent scaffold can be derivitized using saturation mutagenesis, or other PCR-based technique, to create a combinatorial library of structure homologous proteins. Moreover, the putative RNA binding hemisphere of U1A provides a large surface area to envelop the three-dimensional tertiary structure of folded RNAs, specifically RNA hairpins.

As an expansion of our previous work on the semi-designed U1A-TAR RNA binding interface, we looked to develop a higher-throughput platform to screen for artificial RRMs that bind a defined target RNA. This should allow us to affinity mature putative RNA-binding sub-

domains on the U1A scaffold toward rapid generation of new RRMs with improved RNA target affinity and selectivity profiles. First, we screened a β2-β3 loop library to identify an appropriate shape complementarity between the RRM and RNA target. This is followed by parallel screening of β1-α1 and C-terminal helix libraries to distill RRMs with improved contacts with the bound RNA target.

Many high-throughput screening options are available for molecular evolution of proteins, including microarrays[21], phage display[17,22], mRNA / ribosome display[23–25], and yeast display[26–29]. Microarrays are not optimal due to their high cost and instrumentation requirements. Similarly, mRNA / ribosome display are specialized methods that are operationally challenging. Phage display is a common method for molecular evolution of protein function and has been used to identify U1A variants with altered RNA binding properties[11,17]. However, yeast display **Figure 5.2** affords compelling benefits for our purposes in comparison to other methods, including phage display.

First and foremost, the yeast display protocol is operationally simple compared to other techniques. Methods for genetic manipulation of yeast are robust and well detailed in the literature[26,30,31]. Also, the analysis method employs Fluorescence Activated Cell Sorting[32–34] (FACS) **Figure 1.5**, which is very high-throughput and requires few bench top manipulations. Consequently, iterative diversification and screening steps are easily executed. Second, the binding signal in yeast display correlates with the affinity of the measured binding interaction, thus quantitative affinity data can be derived directly in the yeast display system (i.e. the "fitness" of a library population can be determined during the screen)[26,35]. Phage display does not afford

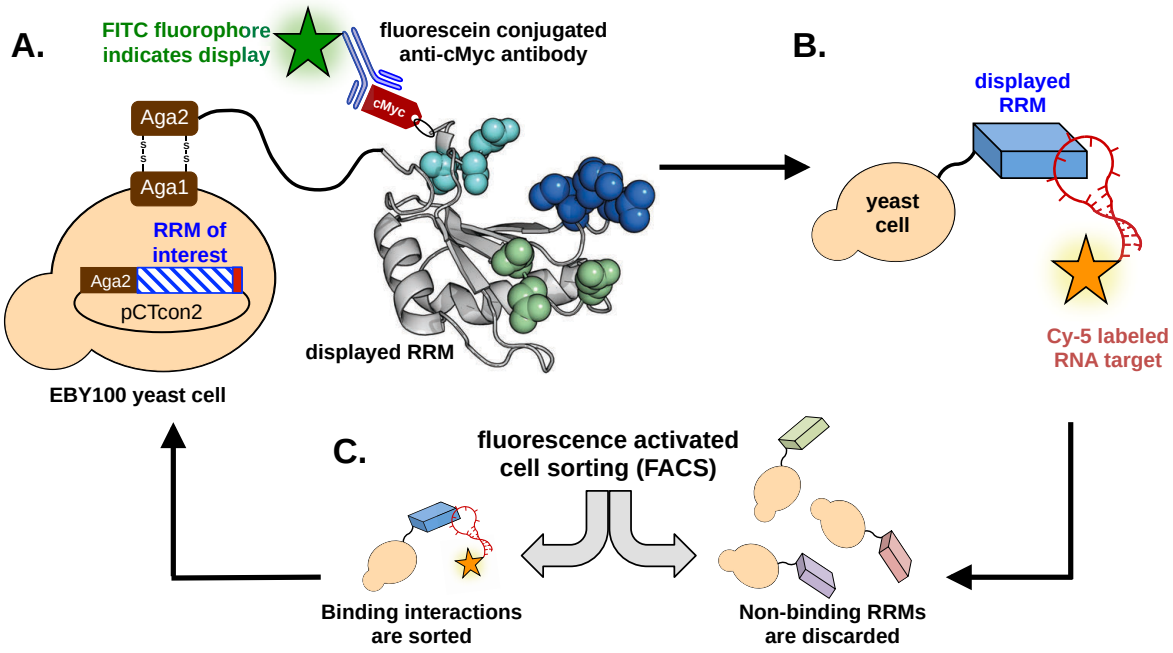


Figure 5.2 A The pCTcon2 plasmid introduced into EBY100 yeast cells encodes an Aga2-RRM chimera. Expression of this fusion leads to display of the RRM on the outside of the yeast cell membrane. A fluorescein conjugated anti-cMyc antibody allows display to be quantified by fluorescence-based flow cytometry. **B** Yeast cells are incubated with 5′-Cy5 labeled target RNA. This labels functional RRMs for fluorescence-based flow cytometry. **C** RRMs that do not bind RNA do not have a Cy5 label and are discarded by fluorescence activated cell sorting (FACS). Cy5 labeled RRMs are collected for subsequent rounds of screening. (PDB code: 1URN)

this luxury.

Lastly, in a yeast display screen ~50,000 copies of a library member are displayed on a single yeast cell[26]. This minimizes false positives, as each positive binding signal represents ~50,000 duplicates of the measured interaction. Additionally, this high copy number results in excellent sensitivity so that binding interactions of up to ~5 μM dissociation constant (K_D) can be observed. This is an ideal starting point for us to begin diversification of the U1A scaffold as the tightest binding variant we identified from our previous studies gave a K_D of 4.1 μM^2 **Table 4.2 entry 2**. Collectively, these qualities make yeast display an attractive option for our purposes.

In the effort described here we randomize residues on putative RNA binding sub-domains in the U1A scaffold and screen ~3×10^7 unique RRMs for improved affinity and selectivity for the HIV-1 TAR RNA hairpin. Our results indicate that using this alternative approach to RNA recognition, we can indeed generate U1A-derived RRMs with excellent affinity and selectivity for a new RNA hairpin target.

5.2 RNA Recognition Motifs Display on Yeast and Bind RNA

Many different classes of proteins have successfully been displayed on the outside of yeast cells[28]. As eukaryotes, yeast express chaperone proteins to assist in proper folding of expressed proteins. As an initial test of yeast display in our hands, we wanted to ensure that both the wt U1A and U1A_E19S scaffolds could be displayed on yeast cells and that displayed RRMs still retained RNA binding activity. Components of the yeast display system were kindly provided by the Dane Wittrup lab, which include a display plasmid, pCTcon2, and an engineered strain of *S.cerevisiae*, EBY100. The EBY100 strain constitutively expresses the membrane bound portion of a yeast cell adhesion receptor, Aga1[27]. To display a protein, the protein of interest is expressed off of the display plasmid as a fusion the extracellular portion of the yeast cell adhesion receptor, Aga2. Aga2 forms two disulfide bonds with the membrane anchored Aga1 protein, tethering the chimera to the outside of the yeast cell membrane.

For our initial test, wt U1A and U1A_E19S were each cloned into the display plasmid as C-terminal fusions to Aga2, with a cMyc tag at the C-terminus of the entire chimera. The new

constructs were separately transformed into EBY100 yeast by electroporation, followed by induction off of the Gal promoter to separately express cMyc tagged Aga2 fusion proteins. After >36 hours of induction, these cells were treated with a fluorescein conjugated anti-cMyc antibody followed by flow cytometry analysis to assess display efficiency. The resulting histogram **Figure 5.3A** shows that >20% of both populations successfully displayed either wt U1A or U1A_E19S, which is consistent with previously reported display efficiencies.

To our knowledge there are no reports of a protein-RNA interaction being monitored by yeast display in the literature. To ensure that yeast display is a viable strategy for monitoring protein-RNA interactions, we treated a separate aliquot of the yeast cells displaying Aga2-wtU1A with varying concentrations of a 5′-fluorescein tagged U1hpII RNA. A fluorescent signal was prominent at 100 nM U1hpII RNA **Figure 5.3B**, suggesting that this protein-RNA binding interaction can in fact be observed by yeast display. Moreover, these data represent, to our knowledge, the first demonstration of a protein-RNA interaction visualized on a yeast cell surface.

5.3 Molding the β2-β3 loop for Shape Complementarity to TAR RNA

Our next step was to mutate residues in putative RNA binding sub-domains on our scaffold with the goal of improving affinity and selectivity for a new RNA hairpin target. For these initial proof of concept studies we chose to begin with the U1A_E19S scaffold, which was previously identified to have improved affinity (K_D = 4.1, ~5-fold improvement over wtU1A

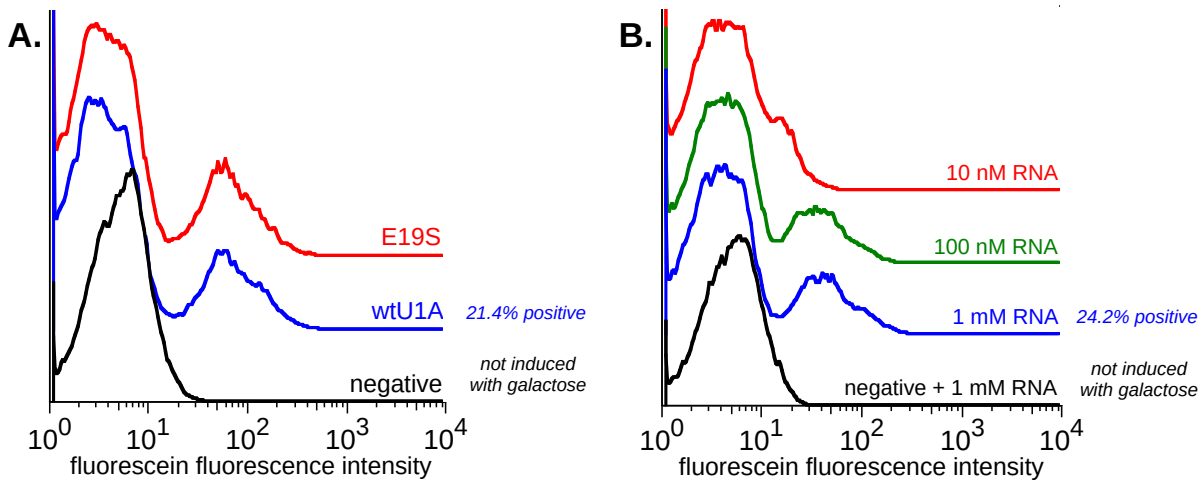


Figure 5.3 A Yeast displaying either cMyc tagged wtU1A or cMyc tagged U1A_E19S were labeled with a fluorescein conjugated anti-cMyc antibody. The histogram demonstrates that both RRMs can be displayed on the outside of the yeast cell membrane. **B** Yeast displaying wtU1A were incubated with varying concentrations of fluorescein tagged U1hpII RNA. The histogram demonstrates that wtU1A displayed on the yeast surface is still functional in binding its cognate RNA.

Table 4.2 entry 2) and great selectivity for the HIV-1 TAR RNA hairpin[2]. Generating new binding reagents for TAR RNA affords us a number of opportunities for downstream evaluation of our approach. Primarily, this RNA hairpin has frequently been the target of other RNA-binding reagents[36–40], which allows us to directly compare our reagents against the state of the art in RNA molecular recognition. Second, assays that measure TAR-dependent HIV-1 mechanisms are widely reported in the literature[41–45], offering a way to evaluate the activity of artificial RRMs in a live cellular environment.

In the native U1A-U1hpII complex, the β2-β3 loop of U1A lodges into the hairpin loop of U1hpII[3,5,16] **Figure 5.1A**. As mentioned, observations from previous studies, and our own work (**Chapter 3**), led us to propose that the size and chemical composition of this loop contributes significantly to the shape complementarity of the RNA binding interface[1]. Therefore, we looked to first randomize this sub-domain by saturation mutagenesis and screen for improved affinity

for TAR RNA hairpin. Through this process, we hoped to identify a new β2-β3 loop identity, or set of identities, that adopt an optimized fit for the TAR RNA hairpin.

Using saturation mutagenesis, a β2-β3 loop library was generated by replacing select loop codons (Ser46, Ser48, Leu49, Lys50, Met51) with NNK codons, where N represents any one of the four nucleotides and K represents guanine or thymine. All twenty amino acids are still potentially encoded by an NNK codon, but two of the three stop codons are inaccessible, thus lowering the frequency of incomplete and potentially non-functional RRMs. Randomization generated a theoretical library size of 3.2×10^6 RRMs encoded as C-terminal fusions to Aga2 in the display plasmid pCTcon2. Transformation into EBY100 yeast gave ~6.6×10^6 colony forming units (cfu).

This number of transformants likely does not exhaustively incorporate the entire theoretical library size, however, this transformation efficiency is excellent relative to other yeast transformation protocols in the literature[26,30]. We acknowledged and accepted this tradeoff of yeast display in return for the benefits if affords, detailed above. Also, given the privileged characteristics of our starting scaffold, we reasoned that an exhaustive sampling of sequence space would not be essential to identify RRMs with improved affinity for our target. Pleased with this library size, we moved forward to screen it for improved affinity.

5.4 Yeast Display Screen of the β2-β3 Loop Library

A yeast culture, encoding the β2-β3 loop library was made to express and display the

Aga2-β2-β3 loop library-cMyc tag chimera protein **Figure 5.2A**. For the initial binding conditions, we incubated these yeast with 10 μM 5′-Cy5 (a small molecule fluorophore with an orthogonal excitation / emission spectrum to fluorescein) labeled TAR RNA **Figure 5.2B**. This relatively high concentration of target RNA essentially allows us to enrich for RRMs that fold properly, do not contain a stop codon in the randomized region, and have some baseline affinity for TAR RNA. Additionally, we were also interested in validating the screening conditions, given the lack of precedent for observing protein-RNA interactions by yeast display.

A universal concern in screening for RNA binding reagents is pulling out molecules with numerous positive charges that non-specifically bind all RNAs. To identify selective TAR RNA binders, we wanted to preempt the selection of RRMs with multiple positively charged residues that ubiquitously bind RNA molecules. Thus, we included 5 μM *E.coli* tRNAs in the incubation with the yeast and the Cy5 labeled RNA. The tRNAs function as a good approximation of non-specific competitor RNAs[46,47]. First, tRNAs are ubiquitously present in the cytosol, so artificial RRMs resulting from our screen would need to orthogonally bind the target RNA in the presence of tRNAs to be viable reagents *in vivo*. Second, tRNAs contain a number of secondary structures, including hairpins. We also included a fluorescein conjugated anti-cMyc antibody into the incubation to measure display efficiency. The binding mixture was allowed to incubate at room temperature for 1 hour with agitation.

Unbound RNA and antibody were washed away before analyzing the yeast cells by flow cytometry. The orthogonal excitation / emission spectra of fluorescein and Cy5 allows concomitant analysis of display and RNA binding activity, as verified by controls **Figure 5.4B**.

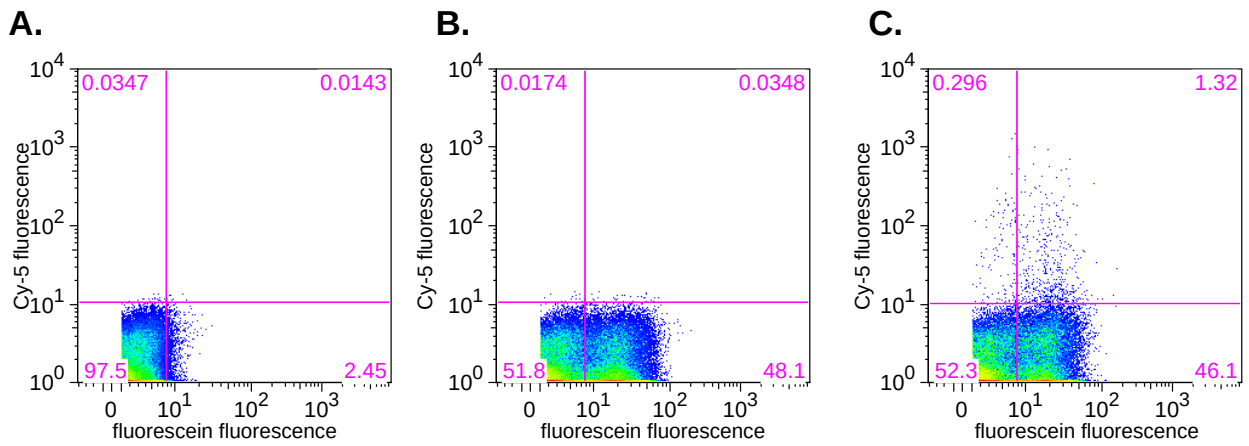


Figure 5.4 A Negative control for yeast display. Cells were not induced to express and display an RRM. **B** Cells induced to express a cMyc-labeled RRM were only labeled with a fluorescein conjugated anti-cMyc antibody to demonstrate that the fluorescein channel does not overlap with the Cy-5 channel. **C** The b2-b3 loop library RRMs displayed on yeast cells were treated with both antibody and 5′-Cy5 tagged TAR hairpin RNA. A representative population is shown. In total, 4.68×10^7 events were observed and 1.32×10^6 double positive cells were collected.

A scatter plot representing each yeast cell in the population is generated with the indication of display, fluorescein fluorescence, plotted on the x-axis and binding indication, Cy5 fluorescence, plotted on the y-axis, aligning with the convention of yeast display protocols[26]. The bottom right quadrant represents yeast cells displaying a library member that does not bind the target RNA. We were pleased to find that >2% of the 4.68×10^7 events observed were double positive for both fluorophores, showing up in the top right quadrant of **Figure 5.4C**. The double positive cells (1.32×10^6 events) were collected into the appropriate growth media using fluorescence activated cell sorting (FACS) **Figure 1.5**. These collected cells represent the first generation of TAR RNA-binding RRMs, or TBR_1G. Display plasmids encoding the identity of individual TBR_1G members were retrieved from a small sample of the collected cells and sequenced. Sequencing data showed unique β2-β3 loops in all members analyzed, suggesting a combinatorial population is maintained after the first sort.

The population of yeast cells harboring TBR_1G members was cultured and then induced again for subsequent sorting steps. With a now enriched population of RRMs, we increased the stringency of binding conditions over the next two sorts to isolate better TAR RNA hairpin binders. The second and third sorts were performed in sequence without any intermediate diversification. The 5′-Cy5 labeled TAR RNA concentration in the binding mixture was decreased to 1.0 μM and 500 nM in the second and third sorts, respectively. In both sorts, the incubation time was also decreased from 60 to 30 minutes. The tRNA concentration remained the same at 5 μM. In both sorts ~3×10^7 events were observed. In the second sort 6.5×10^5 yeast cells gave a double positive signal and were collected. Similarly, 2.48×10^5 yeast cells were collected in the third sort. A small sampling of each sorted population also showed unique sequences in the β2-β3 loop. The TBR_3G members likely contain β2-β3 loops with better shape complementarity or improved contacts for TAR RNA than their predecessors.

5.5 Secondary Sub-Domain Maturation of the Enriched RRM Scaffold

After the third generation, we next looked to introduce more sequence diversity into the enriched population of TBR_3G. Little is known about the TAR RNA-RRM binding interface, given its novelty. Therefore, we turned to the RNA binding features of the parent scaffold, U1A, to inform decisions on where to introduce new sequence diversity on the TBR_3G scaffold. As described above, residues on the β1-α1 loop[17,18] and the C-terminal[12,14] helix contribute significantly to the stability of the native complex through direct contact with loop bases of the

cognate RNA, U1hpII. Again, given the lack of binding data, optimization of either of these spatially segregated sub-domains has an equal likelihood of improving affinity for TAR RNA. Also, it is operationally simple to randomize both sub-domains and screen for improved affinity using yeast display. Binding affinity for TAR RNA correlates with Cy5 fluorescence intensity[26,35], thus allowing facile quantification of library fitness.

Each of the two subdomains, the β1-α1 loop and the C-terminal helix, were separately randomized within the TBR_3G scaffolds using saturation mutagenesis to give two combinatorial RRM libraries that share β2-β3 loop identities **Figure 5.5**. To determine the fitness of each library they were first separately transformed into EBY100 yeast, yielding 3.2×10^7 and 2.1×10^7 cfu for the β1-α1 loop and the C-terminal helix libraries, respectively. We estimate the theoretical library size to be in the same approximate range of the number of cfu, suggesting that together, the new libraries represent ~3×10^7 unique sequences[48,49]. Transformed yeast were then made to express and display their encoded RRM and the two libraries were screened in parallel.

This population of RRMs has eight total residues that are potentially different from the parent scaffold. This prompted us to increase the pressure for TAR RNA selectivity to minimize non-specific binders with multiple positive charges. Therefore, in the fourth generation screening conditions the tRNA concentration was increased by ten-fold, to 50 μM. This tRNA concentration closely mimics to the estimated cytosolic concentration of tRNAs[47]. The additional mutations to the scaffold also increase the potential for more productive contacts with the target RNA. Toward the goal of applying newly identified RRMs to *in vivo* applications

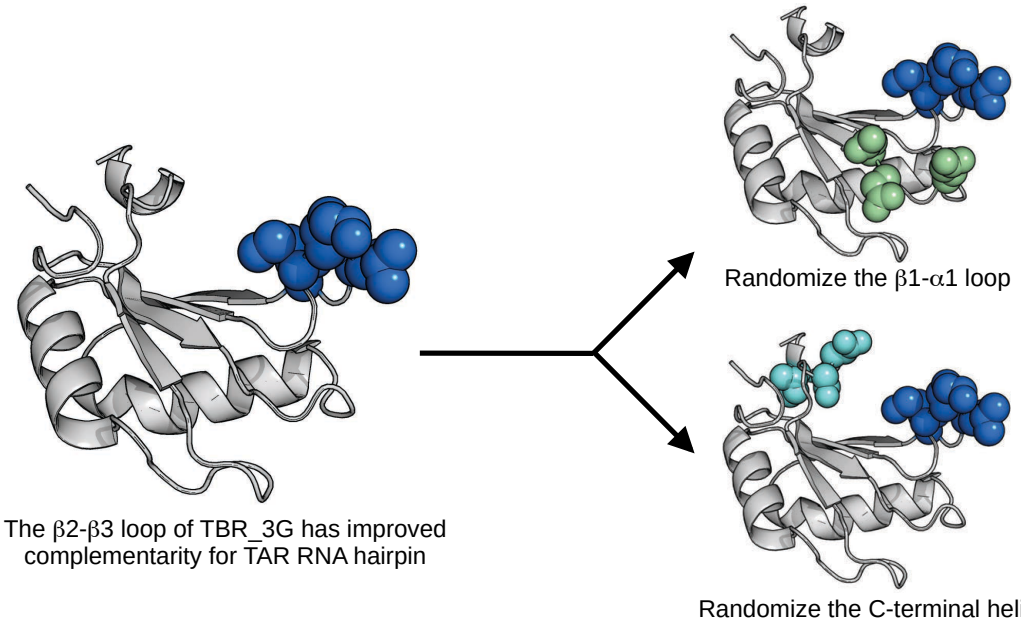


Figure 5.5 After three rounds of sorting the β2-β3 loop library, selected scaffolds were diversified at either the β1-α1 loop or the C-terminal helix to generate two separate libraries. (PDB code: 1URN)

and select for more stable complexes, we increased the temperature of the binding incubation to physiological temperature. The 5′-Cy5 labeled TAR RNA was kept constant relative to the previous sort, at 500 nM.

After incubation and washing, both libraries were sorted sequentially and ~3×10^7 events were observed for each library. The C-terminal helix library gave about 5- fold more double positive events than the β1-α1 loop, suggesting that mutations to this region are significantly more important for TAR RNA recognition **Figure 5.6A** and **B**. The sorted population for both libraries, β1-α1 loop TBR_4G and C-terminal helix TBR_4G, were cultured and a sample from each was sequenced. Sequencing suggests a diverse population in both libraries.

To further compare these two regions, both 4G enriched populations were cultured and induced for another round of sorting. For fifth round binding conditions, the 5′-Cy5 labeled

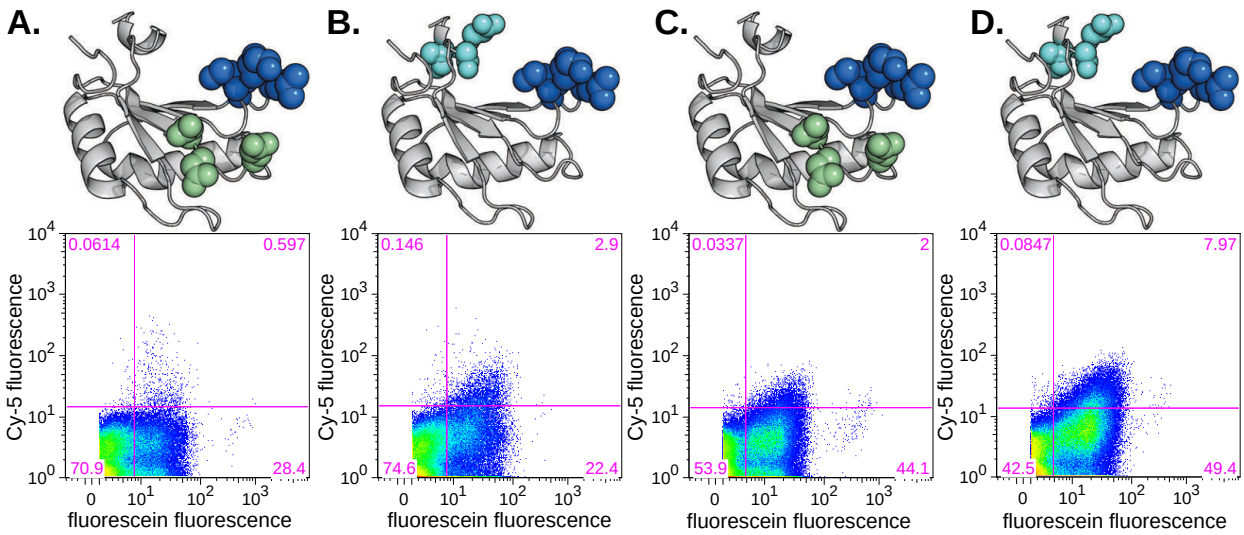


Figure 5.6 A and **B** The fourth round of sorting compares the TAR RNA binding affinity of the β1-α1 loop library and the C-terminal helix library in the presence of 50 μM competitor tRNAs and 500 nM TAR RNA. The C-terminal helix library contains more double positive clones than the β1-α1 loop library. **C** and **D** The fifth round of sorting compares the two libraries in the presence of 50 μM competitor tRNAs and 100 nM TAR RNA. Again, the C-terminal helix library shows more double positive clones. (PDB code: 1URN)

TAR RNA concentration was decreased to 100 nM. All other conditions were kept the same (50 μM tRNA, 37 °C for 30 minutes). After incubation and washing, both populations were sorted by FACS and ~3 × 10^7 events were observed for each. Again the C-terminal helix library gave better indications of binding, showing ~5.7- fold more positive members **Figure 5.6C** and **D**. This strongly suggest that this sorted C-terminal helix population is more fit than the β1-α1 loop population. The β1-α1 loop population was therefore discarded and the more fit population was carried forward as TBR_5G.

In the final sort, we looked to decrease the target RNA concentration into the low nanomolar range (~10^{-8}). For the sixth round binding conditions, we prepared three separate incubations with varying concentrations of 5′-Cy5 labeled TAR RNA: 50 nM, 25 nM, and 10 nM. All other conditions remained the same (50 μM tRNA, 37 °C for 30 minutes). After

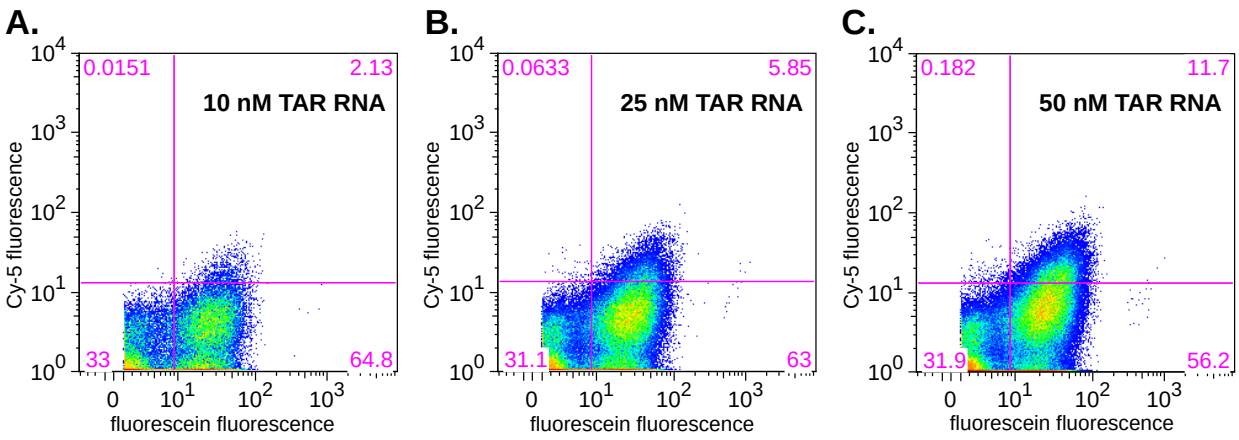


Figure 5.7 Before the sixth round of sorting, three separate concentrations of RNA were analyzed. **A** At 10 nM, we still observed binding in 2% of the population. These double positive cells were sorted to generate TBR_6G. **B** and **C** The number of double positive cells increases linearly with the RNA concentration.

incubation and washing, all three were first analyzed by flow cytometry. Amazingly, 2.13 % of the population appeared double positive in 10 nM of target RNA **Figure 5.7A**. From these double positive cells, 2.5×10^5 cells were collected and cultured. Samples with 25 nM and 50 nM target RNA showed linear increases in double positive cells, as expected **Figure 5.7B** and **C**.

5.6 Binding Characterization of Individual TBR_6G Members

A valuable feature of yeast display is that in addition to functioning as a high-throughput screening method, it also provides quantitative binding data[26,50]. Interactions can therefore be characterized within the system they were identified. Six clones from TBR_6G were sequenced and individually transformed back into EBY100 yeast. Homogeneous samples of yeast, each harboring a unique TBR_6G, were prepared and incubated in conditions identical to the sixth round sort (10 nM 5′-Cy5 labeled TAR RNA, 50 μM tRNA, 37 °C for 30 minutes). After

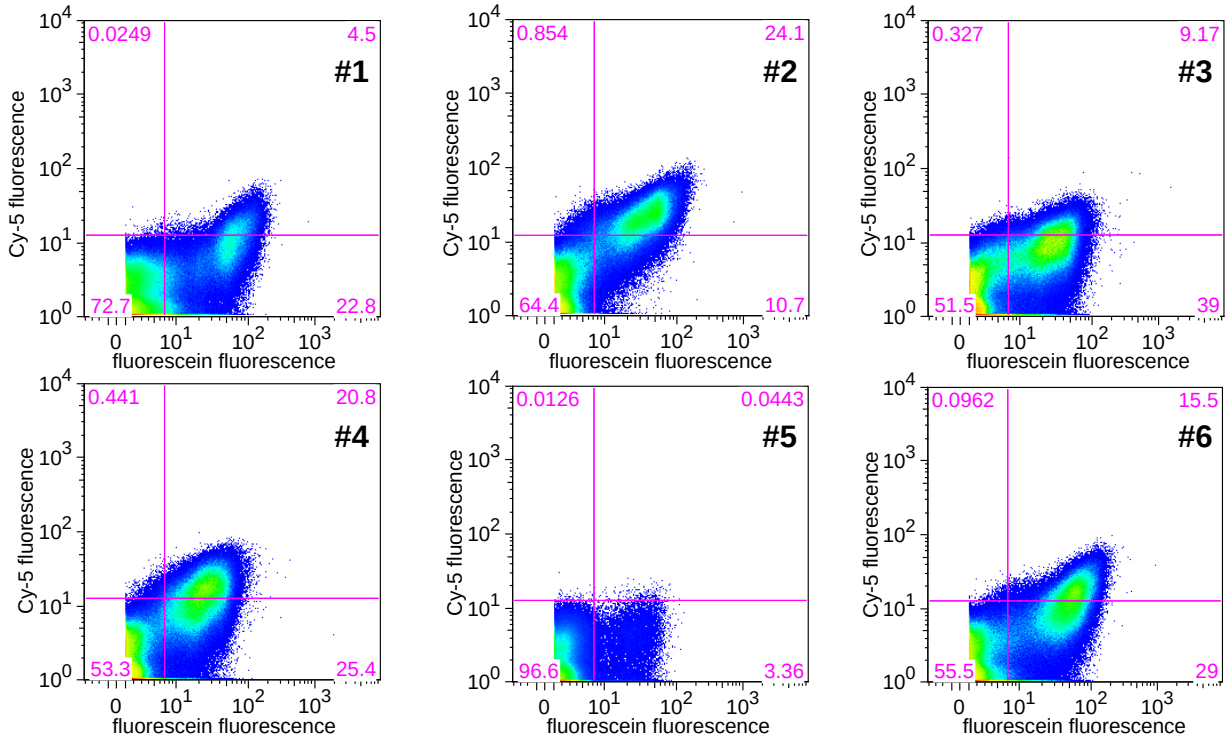


Figure 5.8 Six randomly selected members of TBR_6G were individually analyzed by yeast display. Clone #5 does not give an appreciable binding signal, however all other clones demonstrate binding to TAR RNA.

incubation and washing, samples were analyzed by flow cytometry.

Of the six TBR_6G analyzed, one appeared negative. The other five showed varying levels of % double positive cells, from 4.5% to 24.1% **Figure 5.8**. This is compelling evidence that displayed RRMs bind to 5′-Cy5 labeled TAR RNA with mid- to low-nM K_D. Robust protocols in the literature show that a K_D can be determined through flow cytomometric analysis of a displayed protein[26,35]. To quantify this interaction, the two of best clones were again prepared for flow cytometry analysis with varying concentrations of 5′-Cy5 labeled TAR RNA. Mean Cy5 fluorescence was plotted against RNA concentration and the data was fit to the Hill equation on a single binding isotherm. This showed K_D's of 242.6 (±23.7) nM and 80.3 (± 8.6) nM for clones TBR_6G #2 and TBR_6G #6, respectively **Figure 5.9**. These preliminary data

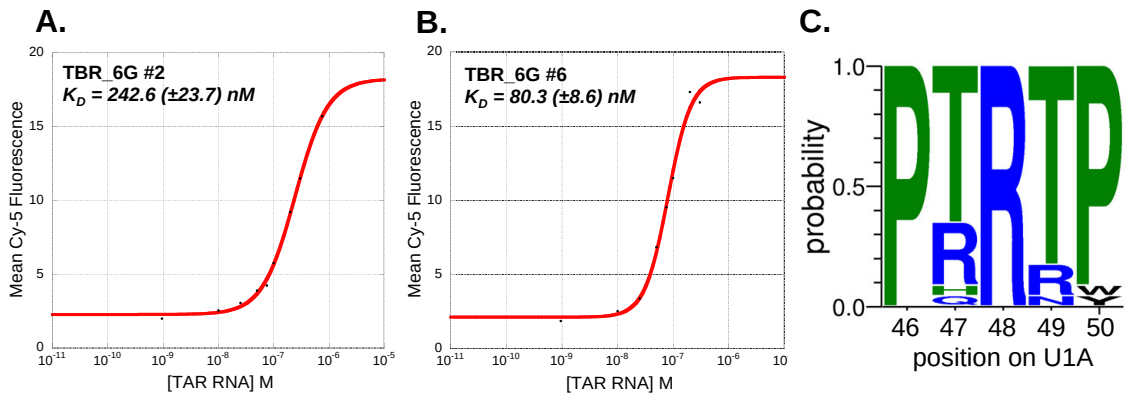


Figure 5.9 Binding data for individual clones as determined by yeast display. The displayed RRMs were kept at a minimum concentration. Concentration of 5′-Cy5 tagged TAR RNA was varied from 500 nM to 1 nM.

represent ~ 17- and ~50-fold improvements in affinity relative to the starting scaffold, U1A_E19S (K_D = 4.1 μM). This is a remarkable observation in comparison to other designed ligands for RNA recognition[39,51–53,54]. The synthetic RRMs reported here bind a folded RNA target with excellent affinity and selectivity, and do so in the presence of physiologically relevant concentrations of non-specific tRNA competitors. With respect to the identity of these synthetic RRMs we can begin to see a consensus sequence emerge in the β2-β3 loop region **Figure 5.9c**. Collectively, these data suggest that, as anticipated, this workflow has allowed us to distill synthetic RRMs that are specifically suited for recognition of the target RNA. Further efforts to characterize these binding interactions by isothermal titration calorimetry and surface plasmon resonance are ongoing and will be reported in due course.

5.7 Conclusions

We have previously shown that the RNA binding specificity of the U1A scaffold can be

altered through surface mutations to putative RNA binding sub-domains[2,8,10,17,18]. The efforts described here translate our initial observations into a semi-design platform for high-throughput identification of artificial RRMs that bind a distinct folded RNA target. In doing so, we have answered some fundamental questions regarding the potential limits and viability of this approach. Here we have demonstrated a level of control over RRM recognition profiles by directing new mutations to improve affinity for a desired RNA target.

We first screened random permutations of the β2-β3 loop to find better shape complementarity for the TAR RNA hairpin. Over three generations of screening, this yielded a population of RRM scaffolds enriched for TAR RNA recognition. Within these scaffolds we separately randomized two spatially defined putative RNA binding sub-domains, the β1-α1 loop and the C-terminal helix. This generated two independent RRM libraries with shared β2-β3 loop identities. Screening these libraries in parallel identified the C-terminal helix as the more beneficial region to modify for increased TAR RNA affinity and selectivity. Preliminary binding data indicates that newly identified artificial RRMs bind to TAR RNA with mid- to low-nM K_D's. The success of this screen provides strong evidence in support of our semi-design approach to molecular recognition of folded RNAs.

The strategy outlined here is, to our knowledge, the first reported example of selecting RNA binding proteins by yeast display. The screen inherently selects for folding stability, target affinity, and target selectivity simultaneously. Additionally, deconvolution and quantitation of individual library members is operationally simple. This protocol will likely be broadly applicable to a range of RNA hairpin targets. Applying this approach can essentially remold most

of the RNA binding hemisphere of the RRM, suggesting that other folded RNA structures that do not form hairpin tertiary structures may also be amendable as targets for this screen. We look forward to the new basic science reagents and potential therapeutic drug leads that result from this technology.

5.8 Methods

Library cloning Specific sub-domains described in this work were randomized using a modified saturation mutagenesis protocol. Cloning of the β2-β3 loop is provided as a specific example. A portion of the U1A gene, encoding C-terminal residues starting with Arg52, was amplified by PCR (Q5 DNA polymerase). The reverse primer encodes a BamHI flanking site on the 5′ end. The 5′ flanking sequence of the forward primer encodes NheI and BsaI cut sites. BsaI is a type II restriction enzyme that cuts two bases away from its non-palindromic restriction site to create a four base overhang that is completely independent of its restriction site. The resulting PCR amplicon was agarose gel purified and then digested with NheI and BamHI restriction enzymes (NEB), followed by purification with an E.Z.N.A extraction kit (Omega). Similarly, the yeast display plasmid was digested with identical restriction enzymes, treated with calf intestine phosphatase (CIP) (New England Biolabs) and purified from an agarose gel. The insert and vector were ligated together using Quick Ligase (New England Biolabs) and transformed into chemically competent 5-alpha *E.coli* (New England Biolabs) according to the manufacturers instructions. Rescued cells were plated on LB ampicillin agar plates and incubated overnight at

37 °C. Surviving colonies were cultured and plasmid DNA was isolated with an E.Z.N.A. Plasmid Mini Kit I (Omega). Proper cloning was verified by DNA sequencing (Genewiz). The resulting construct was named β2-β3 library-receiving vector.

A saturation mutagenesis reverse primer was designed to encode, from 5′ to 3′, a BsaI recognition site, two nucleotides, MNN codons (the reverse compliment of NNK) to replace β2-β3 loop residues 48-51, the codon for Arg47, another MNN codon for position 46, and an annealing sequence. In the sense NNK codons, N represents any one of the four natural nucleotides and K represents either guanine or thymine. The annealing sequence encodes 24 nucleotides that are complimentary to the DNA sequence directly upstream of the β2-β3 loop. This annealing sequence allows this oligo to function as a reverse primer, in combination with a U1A forward primer with an NheI site, in PCR to amplify the N-portion of U1A. The resulting amplicon encodes, from 5′ to 3′ on the sense strand, an NheI site, U1A residues 1 through 45, an NNK codon, the Arg47 codon, four NNK codons, two nucleotides, and a BsaI site.

The amplicon containing randomized codons in the β2-β3 loop, and the β2-β3 library-receiving vector were separately digested with NheI and BsaI, the plasmid treated with CIP, and both purified. The library insert was ligated into 600 ng of library receiving vector at an insert to vector ratio of 10 to 1, spread over four reactions. The reactions were combined and cleaned via three sequential phenol chloroform extractions, followed by three chloroform extractions. DNA was then ethanol precipitated and resuspended in ddH2O. This DNA was suitable for transformation into *E.coli* to verify the library diversity of the newly created construct.

After verification, this plasmid was used as a template to amplify a homologous recombination insert encoding, from 5′ to 3′, forty bases of the display plasmid, pCTcon2, upstream of the U1A gene, the U1A gene with a randomized β2-β3 loop region, and forty bases of pCTcon2 downstream of the U1A gene. The vector was prepared for homologous recombination by digestion with NheI and BamHI restriction enzymes (New England Biolabs), followed by agarose gel purification. This vector (1,000 ng) and the homologous recombination insert (4000 ng) were mixed and ethanol precipitated together to give a purified DNA pellet suitable for yeast transformation. The EBY100 yeast were prepared for electroporation exactly as described in the Nature Protocols paper, *Isolating and engineering human antibodies using yeast surface display*[26]. Following rescue of the yeast, serial dilutions were plated to estimate library size.

RNA Binding Interactions Yeast cells harboring the appropriate RRM-pCTcon2 plasmid were made to induce and display their RRM by the protocol in *Isolating and engineering human antibodies using yeast surface display*[26], exactly. After >36 hours of induction at 30 °C, cell density was estimated (~1.0×10^7 cells / mL = 1.0 Abs OD_{600}), and ~5×10^7 cells were spun down at 12,000 RPM for ~10 seconds. Cells were resuspended in a total volume of 500 μL containing: 30 units of RNasin RNase inhibitor (Promega), the indicated concentration of 5′-Cy5 labeled TAR RNA (Integrated DNA Technologies), the indicated concentration of tRNAs from *E.coli* (Sigma-Aldrich), and a FITC conjugated anti-cMyc antibody (Abcam). Reactions proceeded for the indicated time at the indicated temperature with agitation.

To prepare samples for flow cytometry, reactions were spun down at 12,000 RPM for ~10

seconds and the supernatant discarded. Cells were resuspended in 500 μL ice cold PBS-BSA and immediately spun down again. The supernatant was removed and pelleted cells were stored on ice in the dark until flow cytometry analysis. Negative controls include yeast not treated with RNA or antibody, yeast harboring the RRM-pCTcon2 plasmid of interest but were not induced that were then treated with RNA and FITC conjugated antibody, and yeast cells displaying the RRM population of interest that is only labeled with FITC conjugated antibody.

Before analysis, cells were resuspended in 500 μL of ice cold PBS BSA and immediately loaded onto the flow cytometry instrument, a Flow Cytometer and High Speed Cell Sorter using a solid-state iCyt 635 nm laser - MoFlo (Dako Colorado, Inc.). Cells were sorted with a flow rate of 10 to 12 thousand cells per second into SD-CAA supplemented with penicillin streptomycin. The pCTcon2 plasmid of sorted yeast cells was retrieved using a Zymoprep yeast plasmid miniprep kit (Zymoresearch). Plasmid DNA was then transformed into *E.coli* so that individual clones could be selected and sequenced.

Determination of K_D's Binding affinities were determined using the protocol from *Isolating and engineering human antibodies using yeast surface display*[26]. TAR RNA concentrations were varied from 1 nM to 500 nM. TAR RNA concentration was plotted against mean Cy5 fluorescence for each sample and fit to the Hill equation for single binding isotherm.

5.9 Proteins Used In This Work

U1A

nucleic acid sequence

ATGGCCCAGGTGCAGCTGCAGGTCGACATGGCAGTTCCCGAGACGCGTCCTAAC
CACACTATTTATATCAACAACCTCAATGAGAAGATCAAGAAGGATGAGCTCAAAA
AGTCCCTGTACGCCATCTTCTCCCAGTTTGGCCAGATCCTGGATATCCTGGTAT
CACGGAGCCTGAAGATGAGGGGCCAAGCTTTTGTCATCTTCAAGGAGGTCTCGA
GCGCCACCAACGCCCTGCGCTCCATGCAGGGTTACCCTTTCTATGACAAACCTA
TGCGTATCCAGTATGCGCGCACCGACTCAGATATCATTGCCAAGATGAAAGGCA
CCTTCGGATCGGTCGACTCTAGAGGATCCCCGGTTGCGGCCGCACATCATCACC
ATCATCACGTGGCCGCAGAACAAAAACTCATCTCAGAAGAGGATCTGAATGGGG
CCGCATAG

amino acid sequence

MAVPETRPNHTIYINNLNEKIKKDELKKSLYAIFSQFGQILDILVSRSLKMRGQAFVIFKE
VSSATNALRSMQGYPFYDKPMRIQYARTDSDIIAKMKGTFGSVDSRGSPVAAAHHHHH
HVAAEQKLISEEDLNGAA

U1A_E19S

nucleic acid sequence

ATGGCAGTTCCCGAGACGCGTCCTAACCACACTATTTATATCAACAACCTCAATT
CGAAGATCAAGAAGGATGAGCTCAAAAAGTCCCTGTACGCCATCTTCTCCCAGTT
TGGCCAGATCCTGGATATCCTGGTATCACGGAGCCTGAAGATGAGGGGCCAAGC
TTTTGTCATCTTCAAGGAGGTCTCGAGCGCCACCAACGCCCTGCGCTCCATGCA
GGGTTACCCTTTCTATGACAAACCTATGCGTATCCAGTATGCGCGCACCGACTC
AGATATCATTGCCAAGATGAAAGGCACCTTCGGATCGGTCGACTCTAGAGGTTC
CCCGGTTGCGGCCGCACATCATCACCATCATCACGTGGCCGCAGAACAAAAACT
CATCTCAGAAGAGGATCTGAATGGGGCCGCA

amino acid sequence

MAVPETRPNHTIYINNLNSKIKKDELKKSLYAIFSQFGQILDILVSRSLKMRGQAFVIFKE
VSSATNALRSMQGYPFYDKPMRIQYARTDSDIIAKMKGTFGSVDSRGSPVAAAHHHHH
HVAAEQKLISEEDLNGAA

5.10 Primers Used in This Work

Primer	Sequence
U1A b2-b3 loop library	5′-TATATGGTCTCGCCCCTMNNMNNMNNMNNCCGMNNTACCAGGATATCCAGGATCTGGCC-3′
U1A b1-a1 loop library	5′-ATATAGGTCTCTTATATCNNKNNKCTCAATNNKAAGATCAAGAAGGATGAGCTCAAAAAG-3′
U1A C terminal helix library	5′-TATATGGTCTCTTGGCMNNGATMNNMNNGTCGGTGCGCGCATACTGGATACG-3′

REFERENCES

(1) Blakeley, B. D.; Shattuck, J.; Coates, M. B.; Tran, E.; Laird-Offringa, I. A.; McNaughton, B. R. *Biochemistry* **2013**, *52*, 4745.

(2) Blakeley, B. D.; McNaughton, B. R. *ACS Chem. Biol.* **2014**.

(3) Oubridge, C.; Ito, N.; Evans, P. R.; Teo, C. H.; Nagai, K. *Nature* **1994**, *372*, 432.

(4) Kranz, J. K.; Hall, K. B. *J. Mol. Biol.* **1999**, *285*, 215.

(5) Allain, F. H.; Howe, P. W.; Neuhaus, D.; Varani, G. *EMBO J.* **1997**, *16*, 5764.

(6) Katsamba, P. S.; Myszka, D. G.; Laird-Offringa, I. A. *J. Biol. Chem.* **2001**, *276*, 21476.

(7) Hall, K. B. *Biochemistry* **1994**, *33*, 10076.

(8) Scherly, D.; Boelens, W.; Dathan, N. A.; van Venrooij, W. J.; Mattaj, I. W. *Nature* **1990**, *345*, 502.

(9) Law, M. J.; Rice, A. J.; Lin, P.; Laird-Offringa, I. *RNA* **2006**, *12*, 1168.

(10) Katsamba, P. S.; Bayramyan, M.; Haworth, I. S.; Myszka, D. G.; Laird-Offringa, I. a. *J. Biol. Chem.* **2002**, *277*, 33267.

(11) Laird-Offringa, I. A.; Belasco, J. G. *Proc. Natl. Acad. Sci.* **1993**, *92*, 11859.

(12) Kranz, J. K.; Hall, K. B. *J. Mol. Biol.* **1998**, *275*, 465.

(13) Shiels, J. C.; Tuite, J. B.; Nolan, S. J.; Baranger, A. M. *Nucleic Acids Res.* **2002**, *30*, 550.

(14) Law, M. J.; Lee, D. S.; Lee, C. S.; Anglim, P. P.; Haworth, I. S.; Laird-Offringa, I. A. *Nucleic Acids Res.* **2013**, *41*, 7092.

(15) Law, M. J.; Chambers, E. J.; Katsamba, P. S.; Haworth, I. S.; Laird-Offringa, I. *Nucleic Acids Res.* **2005**, *33*, 2917.

(16) Jessen, T. H.; Oubridge, C.; Teo, C. H.; Pritchard, C.; Nagai, K. *EMBO J.* **1991**, *10*, 3447.

(17) Yan Fan. *Thesis Dissertation* University of Illinois at Urbana Champaign **2009**.

(18) Nolan, S. J.; Shiels, J. C.; Tuite, J. B.; Cecere, K. L.; Baranger, A. M. *J Am Chem Soc* **1999**, *121*, 8951.

(19) Reetz, M. T.; Carballeira, J. D. *Nat. Protoc.* **2007**, *2*, 891.

(20) Romero, P. A.; Arnold, F. H. *Nat. Rev. Mol. Cell Biol.* **2009**, *10*, 866.

(21) Uttamchandani, M.; Wang, J.; Yao, S. Q. *Mol. Biosyst.* **2006**, *2*, 58.

(22) Kehoe, J. W.; Kay, B. K. *Chem. Rev.* **2005**, *105*, 4056.

(23) Horisawa, K. *Nucleic Acids Res.* **2004**, *32*, e169.

(24) Lipovsek, D.; Plückthun, A. *J. Immunol. Methods* **2004**, *290*, 51.

(25) Ma, Z.; Hartman, M. C. T. In *Antibody Engineering*; Humana Press: New York, NY, 2011; Vol. 805, pp. 367–390.

(26) Chao, G.; Lau, W. L.; Hackel, B. J.; Sazinsky, S. L.; Lippow, S. M.; Wittrup, K. D. *Nat. Protoc.* **2006**, *1*, 755.

(27) Boder, E. T.; Wittrup, K. D. *Nat. Biotechnol.* **1997**, *15*, 553.

(28) Pepper, L. R.; Cho, Y. K.; Boder, E. T.; Shusta, E. V. *Comb. Chem. High Throughput Screen.* **2006**, *11*, 127.

(29) Gai, S. A.; Wittrup, K. D. *Curr. Opin. Struct. Biol.* **2007**, *17*, 467.

(30) Gietz, R. D.; Woods, R. A. *BioTechniques* **2001**, *30*, 816.

(31) Tarassov, K.; Messier, V.; Landry, C. R.; Radinovic, S.; Molina, M. M. S.; Shames, I.; Malitskaya, Y.; Vogel, J.; Bussey, H.; Michnick, S. W. *Science* **2008**, *320*, 1465.

(32) Herzenberg, L. A.; Parks, D.; Sahaf, B.; Perez, O.; Roederer, M.; Herzenberg, L. A. *Clin. Chem.* **2000**, *48*, 1819.

(33) SKLAR, L.; CARTER, M.; EDWARDS, B. *Curr. Opin. Pharmacol.* **2007**, *7*, 527.

(34) Cobb, R. E.; Si, T.; Zhao, H. *Curr. Opin. Chem. Biol.* **2012**, *16*, 285.

(35) Feldhaus, M.; Siegel, R. In *Flow Cytometry Protocols Second Edition*; Hawley, T. S.; Hawley, R. G., Eds.; Humana Press, 2004; Vol. 263, pp. 311–332.

(36) Lalonde, M. S.; Lobritz, M. A.; Ratcliff, A.; Chamanian, M.; Athanassiou, Z.; Tyagi, M.; Wong, J.; Robinson, J. A.; Karn, J.; Varani, G.; Arts, E. J. *PLoS Pathog.* **2011**, *7*, e1002038.

(37) Leeper, T. C.; Athanassiou, Z.; Dias, R. L. A.; Robinson, J. A.; Varani, G. *Biochemistry* **2005**, *44*, 12362.

(38) Athanassiou, Z.; Dias, R. L. A.; Moehle, K.; Dobson, N.; Varani, G.; Robinson, J. A. *J. Am. Chem. Soc.* **2004**, *126*, 6906.

(39) Bardaro, M. F.; Shajani, Z.; Patora-Komisarska, K.; Robinson, J. A.; Varani, G. *Nucleic Acids Res.* **2009**, *37*, 1529.

(40) Chittapragada, M.; Roberts, S.; Ham, Y. W. *Perspect. Medicin. Chem.* **2009**, *3*, 21.

(41) Berkhout, B.; Gatignol, a; Rabson, a B.; Jeang, K. T. *Cell* **1990**, *62*, 757.

(42) Weeks, K. M.; Ampe, C.; Schultz, S. C.; Steitz, T. a; Crothers, D. M. *Science* **1990**, *249*, 1281.

(43) Jeang, K.; Chun, R.; Lin, N. H.; Gatignol, A.; Glabe, C. G.; Fan, H. *J Virol* **1993**, *67*, 6224.

(44) Lalonde, M. S.; Lobritz, M. A.; Ratcliff, A.; Chamanian, M.; Athanassiou, Z.; Tyagi, M.; Wong, J.; Robinson, J. A.; Karn, J.; Varani, G.; Arts, E. J. *PLoS Pathog.* **2011**, *7*, e1002038.

(45) Dingwall, C.; Ernberg, I.; Gait, M. J.; Green, S. M.; Heaphy, S.; Karn, J.; Lowe, a D.; Singh, M.; Skinner, M. a. *EMBO J.* **1990**, *9*, 4145.

(46) Phizicky, E. M.; Hopper, A. K. *Genes Dev.* **2010**, *24*, 1832.

(47) Dong, H.; Nilsson, L.; Kurland, C. G. *J. Mol. Biol.* **1996**, *260*, 649.

(48) Patrick, W. M.; Firth, a. E.; Blackburn, J. M. *Protein Eng. Des. Sel.* **2003**, *16*, 451.

(49) Bosley, A. D.; Ostermeier, M. *Biomol. Eng.* **2005**, *22*, 57.

(50) Feldhaus, M. J.; Siegel, R. W.; Opresko, L. K.; Coleman, J. R.; Feldhaus, J. M. W.; Yeung, Y. A.; Cochran, J. R.; Heinzelman, P.; Colby, D.; Swers, J.; Graff, C.; Wiley, H. S.; Wittrup, K. D. *Nat. Biotechnol.* **2003**, *21*, 163.

(51) Rando, R. R. *Specificity in the Binding of Aminoglycosides to RNA, RNA-Binding Antibiotics* **2001**.

(52) Paul, D. J.; Seedhouse, S. J.; Disney, M. D. *Nucleic Acids Res.* **2009**, *37*, 5894.

(53) Hermann, T. *Biopolymers* **2003**, *70*, 4.

(54) Thomas, J. R.; Hergenrother, P. J. *Chem. Rev.* **2008**, *108*, 1171.