DISSERTATION

AFFINITY MATURATION AND CHARACTERIZATION OF NOVEL BINDERS TO THE HIV-1 TAR ELEMENT BASED ON THE U1A RNA RECOGNITION MOTIF

Submitted by

David W Crawford

Department of Chemistry

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Fall 2018

Doctoral Committee:

Advisor: Brian McNaughton

Christopher Ackerson
Eric Ross
Patricia Bedinger

ABSTRACT

AFFINITY MATURATION AND CHARACTERIZATION OF NOVEL BINDERS TO THE HIV-1 TAR ELEMENT BASED ON THE U1A RNA RECOGNITION MOTIF

The increased understanding of the importance of RNA, both as a carrier of information and as a functional molecule, has led to a greater demand for the ability to target specific RNAs, but the limited chemical diversity of RNA makes this challenging. This thesis documents the use of yeast display to perform affinity maturation for the ability of a protein to bind the TAR element of HIV-1, which is a desirable therapeutic target due to its prominent role in the HIV-1 infection cycle. To accomplish this, we used a "semi-design" strategy—repurposing a natural RNA binding protein to bind a different target—by creating a library based on important binding regions (especially the β2β3 loop) of the U1A RRM. Following selection for TAR binding, a strong consensus sequence in the β2β3 loop emerged. The affinity of certain library members for TAR was measured by ELISA and SPR, and it was determined that the best binder (TBP 6.7) had remarkable affinity (K_D = ~500 pM). This TAR binding protein also proved capable of disrupting the Tat–TAR interaction (necessary for HIV-1 replication) both *in vitro* and in the context of extracellular transcription. Through collaboration, we were able to obtain a co-crystal structure of TBP 6.7 and TAR. This crystal structure showed that the overall structure of TBP 6.7 was largely unchanged from that of U1A, thereby validating our semi-design strategy. We also found that the β2β3 loop played a disproportionately large role in the binding interaction (~⅔ of the buried surface area). The prominence of this region's role in the interaction inspired the creation and characterization of peptide derivatives of the TBP 6.7 β2β3 loop. These β2β3 loop derived peptides maintain affinity for TAR RNA (K_D = ~1.8 μM), and can disrupt Tat/TAR-dependent transcription. Ultimately, the project yielded a novel platform of TAR binding peptides and a crystal structure which will inform future RNA targeting efforts in addition to generating the tightest known binder of TAR.

ACKNOWLEDGEMENTS

I would formally like to acknowledge everyone who provided financial support for the research in this thesis, including the NIH, Novartis, and Colorado State University.

I would also like to thank the faculty of the CSU Chemistry Department, for their classroom instruction and generous open door policies, especially Prof. James Neilson for instilling in me a deep-seated love of the perovskite structure (which is always the answer).

I'd like to thank John Anderson and the entire Wilusz^2 Lab for treating a demanding guest as a dear friend. It enabled my most challenging experimental work, and I will never forget it.

I would like to thank my colleagues in the McNaughton Lab for the camaraderie we shared through the highs and lows of graduate school. I'd like to particularly thank Angeline Ta, without whose friendship and scientific support I never would have completed a PhD.

I'd like to acknowledge my mentor, Dr. Brett Blakeley, for teaching me how to hold a pipette, and my mentees, Patrick Beardslee and Zachary Fleishhacker, for being receptive and creative.

I would next like to thank a group of people I I have never met, but with whom I am honored to share an author line: my collaborators. Most especially, I'd like to thank Dr. Ivan Belashov and Prof. Joseph Wedekind for the PDB file that made me cry with delight.

I am grateful for the long-term support from my committee members: Prof. Pat Bedinger who guided me through my first practical foray into molecular biology so many years ago; Prof. Chris Ackerson who mentored me through my first publishable work; and Prof. Eric Ross who treated a one-time rotation student as his very own during all the time I spent at CSU.

Most especially, I'd like to thank my advisor, Dr. Brian McNaughton for always being open to a scientific discussion, giving me sound research advice, and for opening so many doors for me. Brian helped me do excellent work which speaks for itself, and it speaks even more eloquently and convincingly because he also found the collaborators it needed to become truly great.

Finally, I'd like to express my appreciation for the $1 americano refills at Mugs Coffee Lounge which fueled the creation of this thesis, and Maryann Crawford for proofreading it.

DEDICATION

For all the times that I felt my struggles and failures as a scientist were all that defined me, and all the times you showed me they were not: I dedicate this thesis to my mother, my father, and my sisters. Your love and support has made possible everything I've accomplished.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

1.1 Biological Macromolecule Background

The intertwined fields of Biochemistry, Molecular Biology, and Chemical Biology can be imperfectly generalized as the study of the properties and interactions of four major classes of biological polymers- lipids, carbohydrates, proteins, and nucleic acids. Nearly every genetic survey, every pharmaceutical, every *in vitro* cellular study, and every assay in all these fields are focused on some aspect of chemically classifying one of these four macromolecules and their interactions with each other and small-molecule adjuncts to form pathways. These four classes of molecules have not, however, had equal time and resources devoted to them. The degree to which our understanding of biochemistry is based on giving importance to what we *could* study, and that it is based on what we already know is not always appreciated. Practically, this means that proteins are the most studied molecule, nucleic acids a comfortable second, and carbohydrates and lipids far behind

To be fair, proteins account for a vast array of cellular function ranging from catalysis to structure, but the original conception of the other macromolecule classes as mere facilitators of protein chemistry was reductive; Wide-ranging their function may be, but proteins do not tell the whole story of cellular chemistry. Of the non-protein macromolecules, Ribonucleic acid (RNA) especially has emerged as a material responsible for functional entities ill-defined by the old Central Dogma. New perspectives are needed to conceptualize the remarkable scope and dynamism of RNA's cellular roles, and new tools will need to be developed to study it.

The work described in this thesis advances two goals. The first goal, which sees significant progress represented here, is the development and characterization of possible effector molecules for a single functional RNA of biochemical interest—the Trans-activating Response element of HIV-1 (TAR). The second, more abstract goal, for which this work represents a small

but important step, is progress toward the eventual ability to develop binders for arbitrary and specific functional RNAs.

In the service of understanding these two aims, and why they are only recently coming into focus, let us take a look back at how our understanding of biochemistry has evolved, how the path this understanding took has informed our assumptions, and how this has necessarily resulted in limitations. By understanding this history, we can recalibrate our goals and build new tools to achieve them.

1.1.1 Proto-Biochemistry was Focused on Proteins

In the 19th century Dutch chemist Gerardus Johannes Mulder came to realize that a vast amount of biological substances could be well-described by a single empirical formula—$C_{400}H_{620}N_{100}O_{120}P_1S_1$. Mulder's correspondent Jöns Jacob Berzelius suggested that this massive class of substances deserved the name "Protein," which roughly means, from its Greek roots, "of first importance" [1, 2]. This lofty designation certainly reflects the priorities of the nascent field of biochemistry. The focus on proteins as the primary enablers of cellular chemistry grew when James Sumner crystallized Urease in 1926 [3], proving what most enzymologists already guessed—that enzymes were proteins (Nobel Prize 1946) [4]. Early structural biology was almost entirely focused on proteins—most notably Linus Pauling's hydrogen-bond based justification for the structure of the protein α-helix and β-sheet secondary structure elements [5] (Nobel Prize 1954).

This focus on proteins was not unwarranted, as Table 1.1A shows that this class of molecules makes up a majority of the dry mass of the cell, and this massive percentage does, in fact, correlate to a wide array of function.

The common catchphrases of the day demonstrate this explicit focus on protein. One famous example, "One Gene, One Enzyme" [8] (soon amended to "One Gene, One Polypeptide) highlights this focus. The "gene" had been formulated purely as an abstraction, and the first, and

Table 1.1: Composition of a Cell A: Breakdown of mammalian cellular dry mass by molecule type [6] **B**: Breakdown of RNA mass by RNA type [7]

A	% of Dry Mass	**B**	% of RNA
Protein	59.31%	rRNA	80%
DNA	0.82%	tRNA	15%
RNA	3.62%	mRNA	5%
Polysaccharides	6.59%	Other Functional RNAs	0.1%
Lipids	16.47%		
Small molecules	13.18%		

for *decades* primary, concrete conceptual connection between the abstract idea of the heritable gene and the gene's effect on observed reality was as an enabler of protein chemistry.

In the early 20th century, the hypothesis that Proteins were somehow the basis for genetic transfer was taken as a *fait acccompli*. How could a molecule as chemically simple as *DNA*, with only four bases, possibly conduct the complicated business of genetic transfer? However, the definitive proof that the nucleic acids were the physical basis of the gene was discovered through observation of the heritability of infectivity in *Streptococcus pneumoniae* by Oswald Avery, Colin MacLeod, and Maclyn McCarty in 1944 [9]. The fact that Avery did not win the Nobel Prize before his death in 1955 is testament to the unwillingness of his contemporaries to accept his (correct) conclusions, and not until an experiment by Hershey using bacteriophage was published in 1952 did the conclusion that nucleic acids were the basis for genetics become inescapable [10, 11].

1.2 The Central Dogma

1.2.1 Structure and Function of DNA

The scope and promise of the new field of Molecular Biology snapped into focus in 1953 when James Watson and Francis Crick proposed a structure for the Deoxyribonucleic acid (DNA) polymer [12]. A final phrase in this seminal paper, "It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic

material" uses the newly discovered structure of DNA to bridge the gap between the function it had already been determined to have, and a mechanism for fulfilling that function.

The chemical limitations of the nucleic acids was no longer puzzling, it was a *feature*. A sample of DNA was no longer a hodgepodge of functional groups, it was a *strand*, a *linear sequence* which underpins life itself. There is a reason that the enduring public symbol of Biochemistry and Molecular Biology is the elegant DNA double-helix. From the moment the structure was known, the broad strokes of how the information that *is* life had propagated itself from the misty past, and would continue doing so for the foreseeable future, was obvious. The abstract idea of the "gene" was now "information [13]."

Life itself was as readable as this sentence, once the code was learned.

1.2.2 Formulation of the Central Dogma

The code correlating DNA and protein sequence was indeed cracked in a very few years, and Francis Crick formulated the famous Central Dogma of Molecular Biology in 1958:

"[O]nce 'information' has passed into protein it cannot get out again. In more detail, the transfer of information from nucleic acid to nucleic acid, or from nucleic acid to protein may be possible, but transfer from protein to protein, or from protein to nucleic acid is impossible. Information means here the precise determination of sequence, either of bases in the nucleic acid or of amino acid residues in the protein [14, 15]."

This simplifies well to the "DNA→RNA→Protein," approximation taught in high school biology, but all the possibilities that Crick foresaw are shown in Figure 1.1, and Crick's primary assertion—that sequence-based information only flows from nucleic acid to protein—has been remarkably robust. However, the initial role that RNA was given—uninteresting messenger—was quickly found to be too reductive. RNA was soon known to be responsible for catalyzing peptide bond formation in the ribosome, and for acting as an adapter between mRNA and protein primary sequence. These two roles actually encompass ~95% of the mass of RNA in the

cell (Table 1.1) while the sequences and structures of the RNAs involved in these roles are highly conserved across all phyla of life [6].

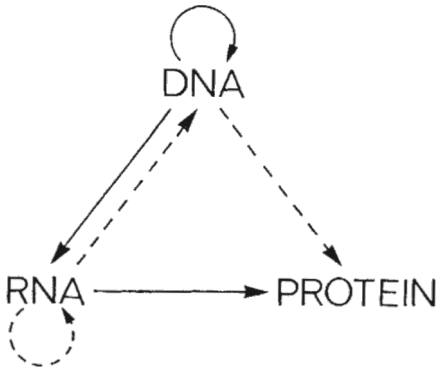


Figure 1.1: The Central Dogma as Formulated by Francis Crick In Francis Crick's own words: "A tentative classification for the present day [1970]. Solid arrows show general transfers; dotted arrows show special transfers. Again, the absent arrows are the undetected transfers specified by the central dogma." Adapted from [15]

1.3 Bioinformatics Era

Counter-intuitive as it may seem, this new focus on DNA actually *increased* the fixation on the fact of proteins as being of "first importance." Once the code was cracked, and it *seemed* clear that DNA didn't (directly) code for anything other than protein, everything else became second-class. If heritable life was "information," and all the chemistry that was coded for was contained in proteins, it follows that the rest of the vast array of chemistry that takes place in the cell can be extrapolated in total if the information in the DNA is sufficiently understood.

The fundamental promise and hope of this formulation was that Molecular Biology is *knowable*. If all function in a cell ultimately derives from DNA, and this DNA is both finite and readable, maybe we could understand, and control *everything*, in a cell. In the 1990s, as the human genome project was nearing completion, this optimism was at its height. Human Genome Project luminary Eric Lander penned a roadmap in 1996 that states "[o]nce all proteins are known, it should be possible to assemble comprehensive 'interaction maps' of genomes. [16]" This assump-

tion that "gene" and "protein" are essentially synonymous, and that a "comprehensive interaction map" *could* be assembled using only the genome are hallmarks of the era.

1.3.1 Gene Editing

The *therapeutic* hope of the bioinformatics era—-that all cell disease states can ultimately be traced back to the genome and corrected there—has been partially borne out. Our increasing knowledge of the genome has indeed allowed us map disorders back to the gene variant which causes them. For instance, Huntington's Disease is understood to be caused by a mutated version of the "Huntingtin" protein which has too many glutamine residues near the N-terminus. We know this because we can read the extra "CAG" repeats in the gene coding for this protein in Huntington's sufferers [17]. We also know that *in cellulo*, correcting the gene fixes the problem [18]. As examples like this demonstrate, there is a great deal of well-placed interest in Gene Therapy [19–21].

1.3.2 Limits

But the rapid DNA-based increase in our knowledge has, until recently, obscured the limits of our current approaches and models. The importance of DNA as the source of the master templates for the proteins in the cell is obvious, but it to call it the "template" or "blueprint" is an overstatement. In hindsight, the belief that gaining a complete understanding of DNA would grant anything like total understanding of cellular processes was extraordinarily naïve. The genome is, more-or-less a parts list, not a manual. Assuming this parts list would be enough to fully understand the dynamic processes that make up up cellular biology is, at a basic level, comparable to assuming that the best way to assemble IKEA furniture is to look at a photograph of a completed item and draw from disorderly piles of every component used by IKEA. Life is not the inevitable result of recorded information, but instead the result of the careful, ordered, limited, and precise *expression* of this information within the pre-existing context of the cell, and within the derived context of a multi-cellular organism.

An example of the power, but also the fundamental limitation, of a DNA informatics approach is the successful synthesis of an artificial genome by Craig Venter's team. This group of researchers synthesized a full genome of *Mycoplasma mycoides* and subsequently successfully transplanting it into a cell, which then grew normally. The achievement is obvious, but so is the limitation. This genome doesn't spontaneously create an organism around it, and would not do so even if transcription/translation machinery and biopolymer building blocks were available. The genome contained the "complete" information for replication, but only if the necessity of an emphM. capriculum cellular environment can be considered free of information [22].

Cellular environment may not follow the clear rules of the DNA code, but it is as important a form of "information" as the easily readable DNA sequences. To take the next steps in biochemical understanding, we need to understand the interaction of every piece of the cell. Arguably the most pervasive and dynamic aspect of that cellular milieu is the seemingly boring RNA.

1.4 The Centrality of RNA to Life

In Molecular Biology as described by the Central Dogma, DNA is the master composer creating a timeless work of genius, Protein is the orchestra expressing that genius, while RNA is relegated to the role of simple amanuensis, making sure the conductor and players have their scores. In Crick's original formulation, RNA dutifully copies down the information contained in the DNA, transmitting it from the nucleus to the parts of the cell that can bring the coded-for proteins into existence, and then degrades without a trace. This is an important job to be sure, but it isn't an apparently complicated one.

However, ever since that boring relegation, RNA continually gets caught doing something interesting. Soon after the formulation of the central dogma, RNA was found to be the adapter between the genetic code and protein synthesis, and to catalyze the synthesis of polypeptides from amino acids within the ribosome. In fact, the ribosomal RNA (rRNA) and transport RNA (tRNA) account for ~95% of the RNA in a mammalian cell (Table 1.1). More recently, it was discovered that RNA doesn't get transcribed from the genome as a perfect copy of a coded gene,

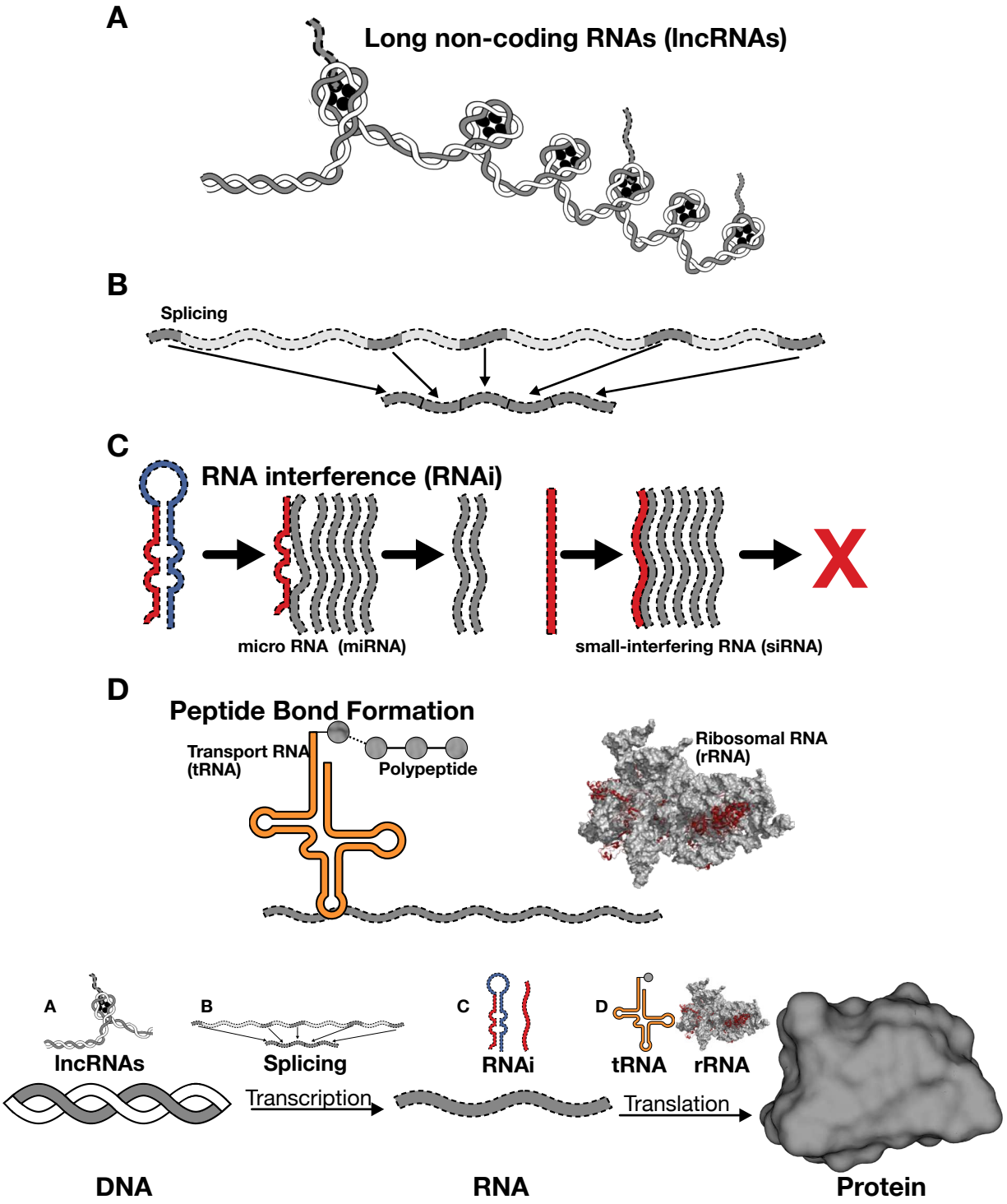


Figure 1.2: RNA Roles in the Central Dogma RNA plays a role in every process described in the Central Dogma, from the **A** long non-coding RNAs that regulate transcription to the **B** post-transcriptional splicing needed to make readable mRNA, to the **C** RNA interference pathways which regulate RNA levels, to the **D** transport and ribosomal RNAs which catalyze polypeptide synthesis.

but needs to be spliced post-transcriptionally into a properly readable form , and that it can even perform this splicing without the aid of protein [23]. More recently still it was discovered that long non-coding RNAs (lncRNAs) regulate gene expression prior to transcription [24, 25], meaning that RNA plays a role in *every* step of the Central Dogma (illustrated in Figure 1.2).

1.4.1 The RNA World Hypothesis

More recent decades have seen the discovery of RNAs that could catalyze reactions, now known as Ribozymes [26], and that the RNA in an RNA/protein complex can act in a catalytic role outside of the Ribosome [27]. This realization that RNA can both carry and replicate information *and* led to the widespread acceptance of the "RNA World" hypothesis, which posits that life as we understand it emerged from self-replicating RNA molecules [28].

Though the RNA World hypothesis is not directly germane to the challenge of targeting any specific RNA, it is *always* worth considering as a universal effector of life. If the RNA World hypothesis is correct, then the chemical trappings of life in all their complexity were pulled into the dance of replication and descent with variation that defines "life" due to their relationship with RNA. It is unsurprising that hardly any processes in a cell that are unaffected by RNA.

1.5 Roles of RNA in Living cells

1.5.1 Dogmatic Roles of RNA

rRNA and tRNA

The vast majority of RNA in the cell (~95%) [6] is either rRNA (~80%) or tRNA (~15%). This is important to consider when designing an RNA binding protein, because these abundant RNAs are *not* viable therapeutic targets. The structures of rRNA and tRNA are complex, varied, and highly conserved among species [7]. This means that any possible therapeutic which has specific activity toward rRNA or tRNA, general activity toward RNA structural elements, or general activity toward RNA, is simply going to be drawn to the ribosome or the tRNA, which every cell needs to survive. Any such therapeutic would be too generally cytotoxic to deserve the name.

RNA as Messenger (mRNA)

The most familiar role of RNA is that of messenger RNA (mRNA), which makes up ~5% of cellular RNA Table 1.1. Messenger RNA is responsible for carrying genetic information (canonically, the primary sequence of proteins) between the DNA in the nucleus which stores this information and the ribosomes in the cytoplasm which express it. Even if this were the only variable class of RNA, mRNA would *still* be a tempting therapeutic target. The genome is more "parts list" than "blueprint", and is identical among all cells in an organism. A human kidney cell and a human brain cell accomplish very different tasks, but the fundamental genetic difference between them is not the information that resides within the nucleus (this is identical), but the information that gets sent *out*.

Messenger RNA is transcribed from the DNA genome. While ~75% of the DNA genome is transcribed into RNA, but only about 2% directly codes for protein. Initial transcripts are known as pre-mRNA [24], and the protein and RNA mediated process known as "splicing" occurs before the eventual mRNA is exported from the nucleus to the cytoplasm. Indeed, without this splicing process, it would not be ribosome-readable. Mis-splicing causes many diseases and disorders, notably certain Muscular Dystrophies (caused by expanded (CUG) repeats), as well as neuron disorders [29]. Specifically targeting such RNAs is an active area of research [29, 30].

1.5.2 RNA Interference (RNAi)

RNAi is the general term for an extensive set of pathways which regulate cellular RNA levels. The study of these pathways, and the RNAs and proteins associated with them, is a field of science in and of itself. Briefly, there are two types of RNAi: microRNA (miRNA) and small interfering RNA (siRNA). Figure 1.3 summarizes the pathways these RNAs participate in.

siRNA

siRNA involves short, single stranded RNAs which are perfectly complementary to an mRNA target. The dsRNA complex formed by the siRNA and the target mRNA activates the protein complex known as Argonaute, which catalytically cleaves any mRNAs complementary to the

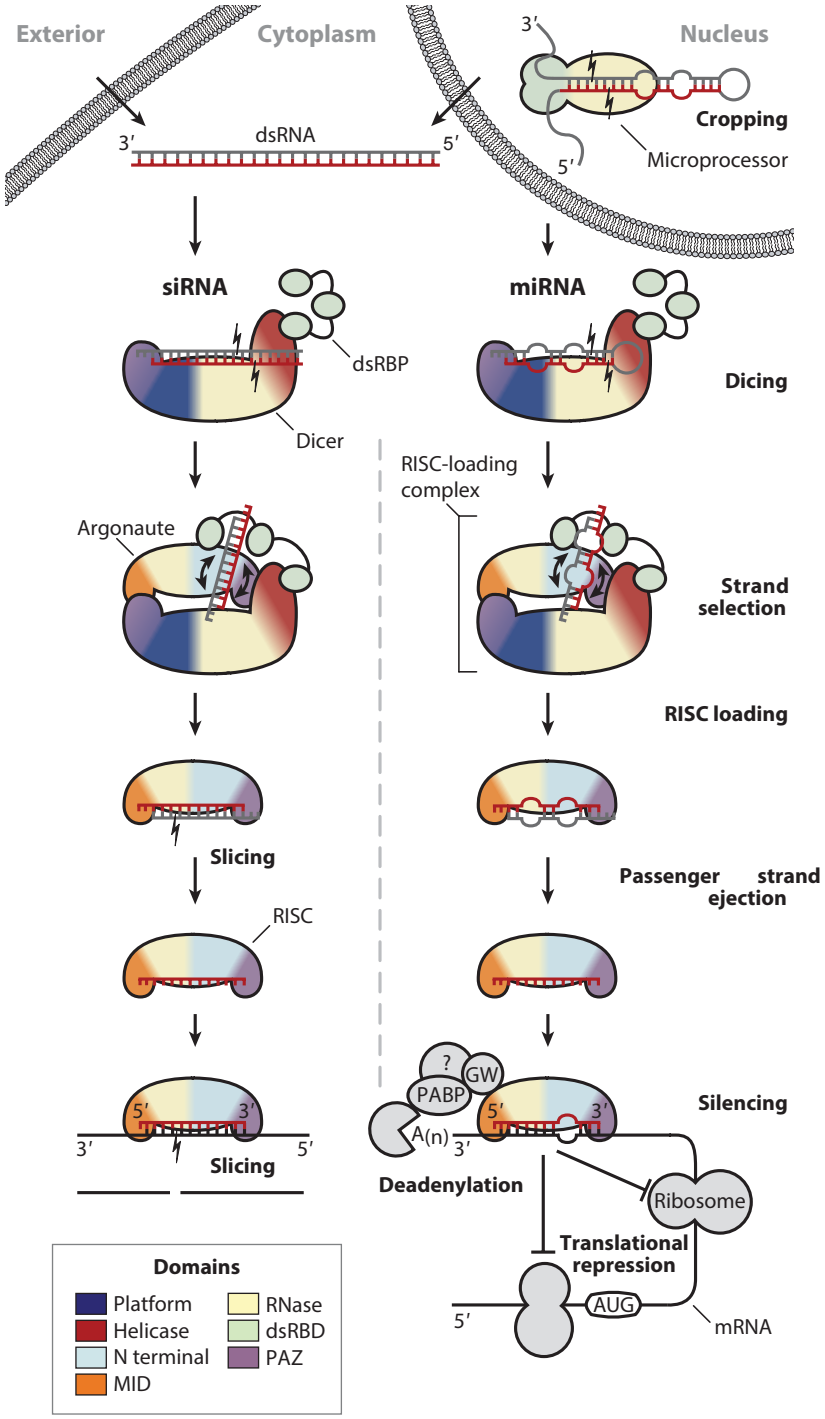


Figure 1.3: Overview of RNA Interference Pathways This figure illustrates the pathways that siRNA and miRNA utilize to regulate gene expression. Though they have different results, both work, generally speaking, through their interaction with the RNA binding protein/nuclease Argonaute. Adapted from [31]

one held in the RISC complex. This is presumably a remnant cellular immune system against dsRNA, since the most reliable source of dsRNA is an infecting virus.

Though the siRNA knockdown pathway is not as robust in mammals as the analogous pathway in plants, when it *does* work, the mechanism of action is catalytic, leading to the targeted mRNA not being translated [31].

miRNA

miRNA uses much of the same machinery, but importantly, they are not perfectly complementary to their targets. They adopt a common conformation involving stretches of complementarity followed by stretches of mismatched "internal loops" (see Section 1.8.2 for a discussion of RNA structural elements). Transcribed miRNAs are not functional, and need processing (shown in Figure 1.3) in order to become active. The active form of an miRNAs is a double stranded stem-loop, and the RNA loaded onto the Argonaute protein which facilitates miRNA regulation is single-stranded. In general, miRNAs have a distinct stem-loop-bulge shape, and a two nucleotide overhang on their 3′ end. A general survey of miRNAs, covering their processing from the transcribed pri-miRNA into functional miRNAs can be seen in Figure 1.4.

These RNAs are involved in *regulation*, rather than the binary knockdown of the vestigial siRNA cellular immune systems, and are enmeshed into the cellular network, with 92% of cellular RNA binding proteins likely involved in miRNA binding [33].

miRNAs are a perfect example of the types of RNAs this thesis hopes to outline general strategies for targeting. They are functional based on both sequence and three-dimensional structure, and many are disease-relevant. For instance, miR-21 is responsible for the regulation of many tumor supressor genes, and its upregulation is associated with many cancers [34]. They are structured, but not distinct enough to be individually targeted by their structure alone. In short, they are the perfect molecular recognition challenge.

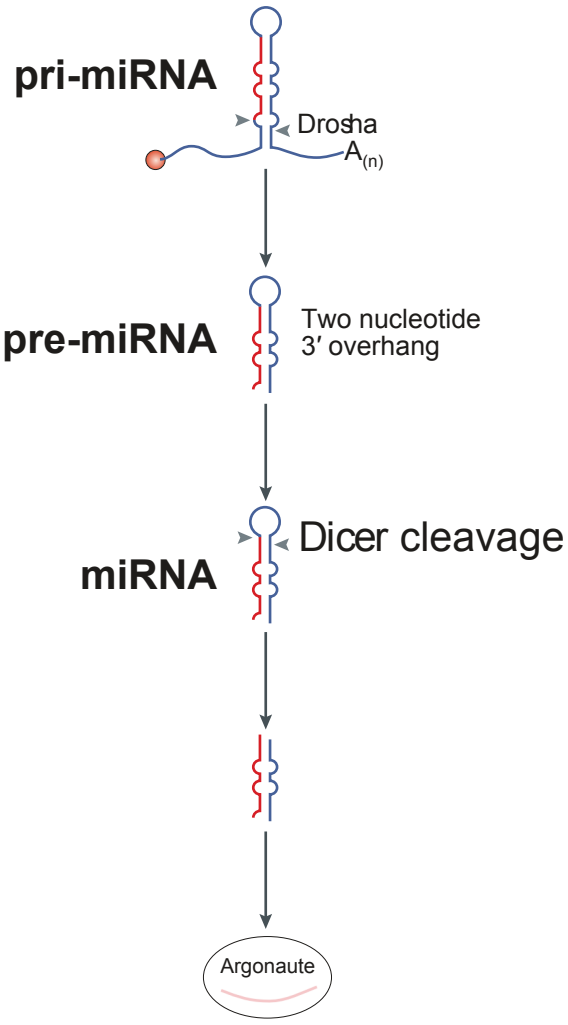


Figure 1.4: micro-RNA Processing Overview The processing steps required between transcription of a pri-miRNA and active regulation. The precise placement of structural elements is notable. Adapted from [32]

.

1.5.3 CRISPR-Cas

Conceptually similar is the famous CRISPR-Cas system, which evolved to enable a prokaryote to store the sequences of viral RNAs into permanent DNA storage is shown in Figure 1.5. The more groundbreaking application has been the adaptation of this bacterial immune response system to write *arbitrary* sequences into DNA storage, and is now most associated with genome editing. Though genome editing is, of course, a DNA modification, it is worth noting that the mechanism of action is predicated upon the Cas9 nuclease binding a guide *RNA*. These RNAs are frequently engineered to be able to bind effector proteins such as fusions to the MS2-coat protein [35]. This use of RNA modifications to CRISPR gRNA is llustrated in Figure 1.5. See Section 1.15.5 for more detail on the MS2 system.

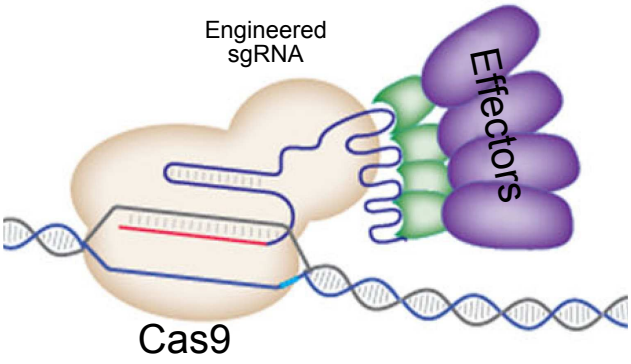


Figure 1.5: CRISPR-Cas9 with RNA Modifications This figure illustrates the CRISPR-Cas9 system, which utilizes a guide RNA (gRNA) as a means of providing a template for editing the genome, and also illustrates a method of manipulating the CRISPR-Cas9 system with a well-understood Protein–RNA interaction—the MS2-MS2 Binding protein pair. Novel synthetic RNA/Protein binding partners would enable further modifications. Adapted from [35]

1.5.4 Viral RNAs

Viruses, minimal as they are, frequently use RNA in a functional manner, and essentially all functional RNAs from viruses that infect human cells can be considered disease-relevant.

Though this thesis focuses nearly exclusively on the HIV TAR element, two additional examples of viral RNAs and their functions are shown in Figure 1.6.

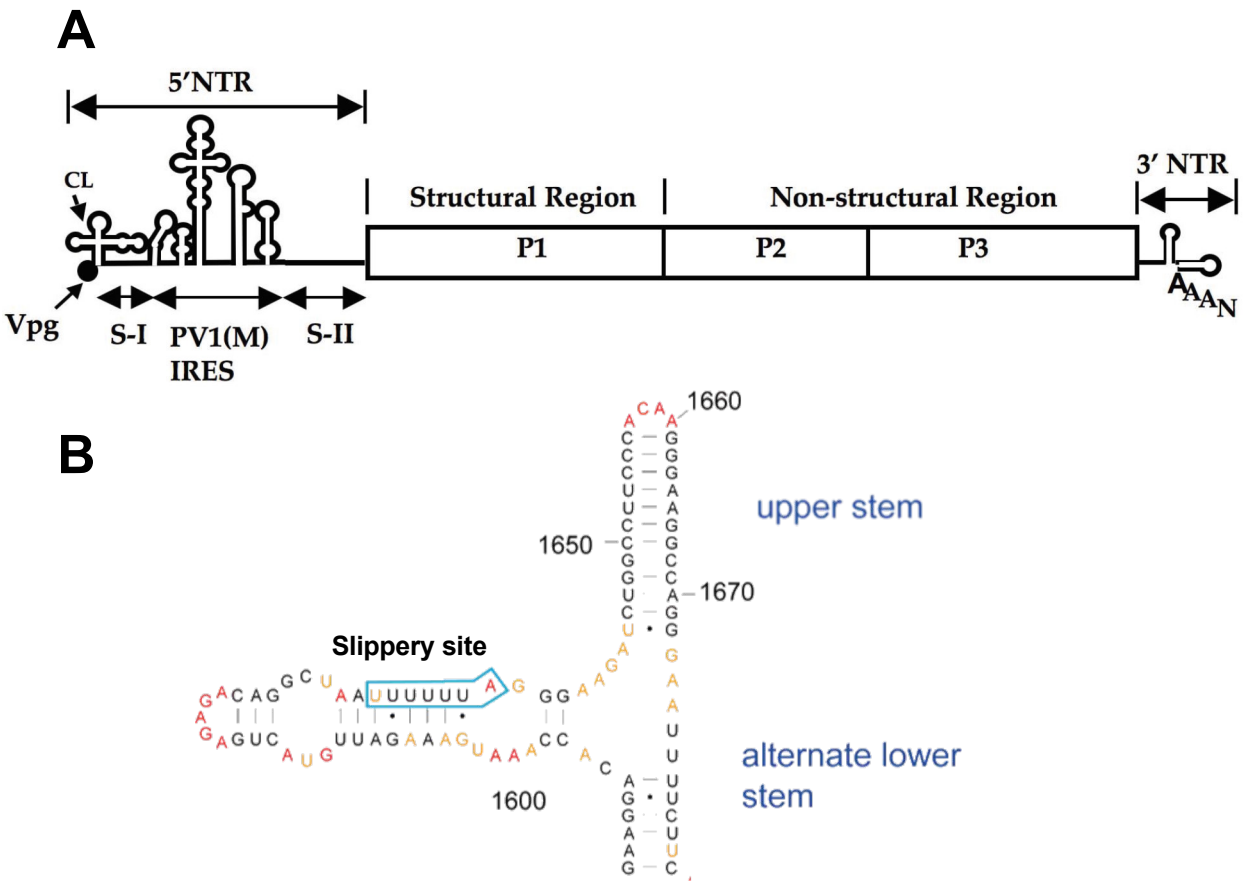


Figure 1.6: Two Examples of Functional Viral RNAs A. The poliovirus Internal Ribosome Entry Site (IRES) Element is required for bypassing the ribosomal requirement for 5′ caps, so the poliovirus ORF can be translated. The structured cloverleaf and V_{pg} elements (which do not code for protein) make up fully ~5% of the poliovirus genome. Adapted from [36]. **B The HIV I Frameshift Element** allows the simultaneous transcription of two overlapping, but frameshifted, open reading frames (*Gag* and *Pol*) into a single polypeptide. Remarkably the ratio of *Gag* to combined *Gag-Pol* transcription is conserved across retroviruses [37]. Adapted from [38]

1.6 HIV I TAR RNA

The *specific* viral RNA target this thesis is most concerned with is the *trans*-activation response element of HIV I (TAR). The TAR region of HIV canonically occupies the first 45 nucleotides of the ~9000 nucleotide HIV I mRNA genome, and is part of the 5′ Untranslated region (5′ UTR)

[39]. TAR plays two important roles in HIV replication (Figure 1.22) which will be discussed in detail shortly, but important to note is that the key facilitator of both roles is a small, structured RNA element occupying positions 17–43 on the HIV genome.

1.6.1 TAR Induces a Transcription Cascade

The first, and most important role TAR plays in the HIV life-cycle, is that of a pseudo-promoter facilitating the replication of HIV mRNA, which in the case of HIV is both template for protein production and packaged genome in the virion. Leaky transcription of the HIV I DNA genome leads to the mRNA transcript of the 5′ UTR of HIV, which base-pairs with the DNA genome. The trans-activator of transcription (Tat) protein from HIV I [40–42], binds the TAR element on the mRNA and recruits the elements from the host cell necessary to transcribe the genome into mRNA with a positive feedback loop [43, 44]. This transcription cascade marks the shift from latent to active HIV infection (illustrated in Figure 1.7).

1.6.2 TAR as pre-miRNA

The second key feature of the TAR element is its ability to act as an anti-apoptotic miRNA, keeping the host T-cell "alive" and manufacturing HIV virions [46–48]. The entire canonical TAR element, as well as a further ~10 nucleotides, are a pre-miRNA, while the mature miRNA is derived from the base-paired region (with mismatches) immediately before and after the structured TAR element (see Figure 1.4 and Figure 1.3 for further details). The TAR element can be seen in the context of the HIV I genome in Figure 1.8.

A key point to note about the two roles of TAR (which are *necessary* for successful HIV I proliferation), is that they are likely less susceptible to mutational evasion as other HIV I targets, such as reverse transcriptase [49–51]. This is due to the fact that, in theory, any mutation which allowed TAR to evade a therapeutic would also require a compensatory mutant to the Tat gene so Tat/TAR-dependent transcription would occur, or would require any mutation to leave the pre-miRNA properties of the TAR element intact.

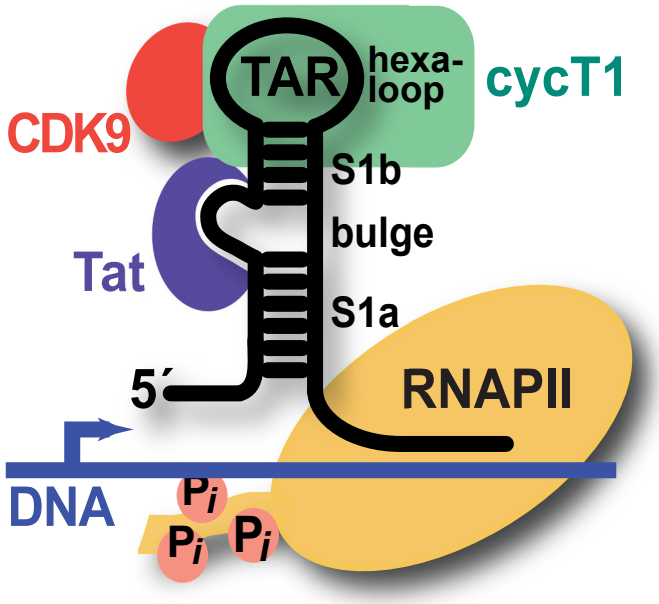


Figure 1.7: Tat–TAR Induced Transcription Cascade The viral transactivation response (TAR) element RNA comprises lower (S1a) and upper (S1b) stems. The positive transcription elongation factor b (p-TEFb) comprising cyclin T1 (green) and CDK9 (red) is recruited to TAR by the HIV-1 protein Tat (purple), which binds the central RNA bulge allowing cyclin T1 to interact with the apical loop. The bound complex stimulates host RNA polymerase II (yellow) by phosphorylation to produce full-length viral transcripts from proviral DNA From: [45]

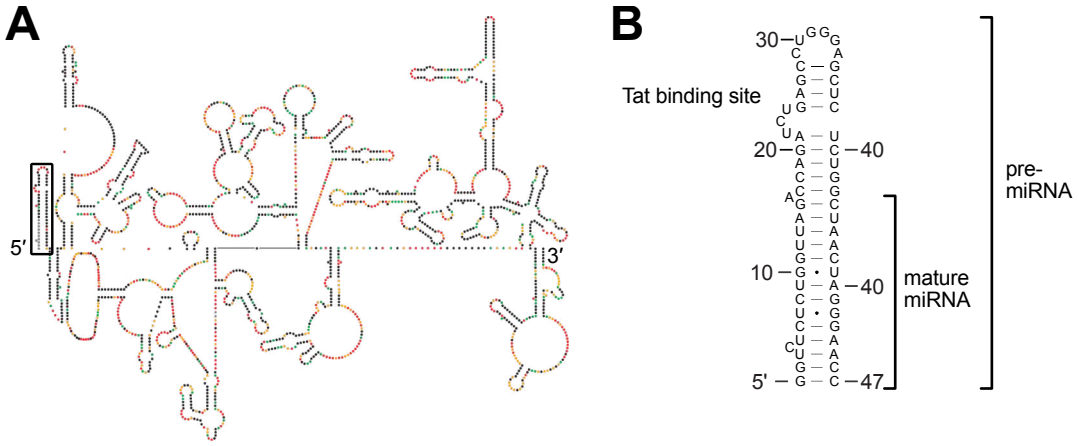


Figure 1.8: The HIV I TAR Element in the Context of the HIV Genome The ~5000 bases on the 5′ end of the HIV I genome, and the secondary structure of same, can be seen in **A**, with the TAR element indicated by a box. **B** shows the regions of the TAR element which engage in the two main functions of TAR: Tat recruitment and miRNA activity.

1.6.3 Relevance of TAR RNA

HIV afflicts over 36 million people worldwide, and no cure (or vaccine) yet exists. Current treatments are effective, but evolving resistance is always a concern, especially since there are only four current drug targets (entry, reverse transcriptase, integrase, and protease) [52, 53].

For reasons outlined above, much effort has gone into generating binders of TAR RNA. These efforts have resulted in TAR-binding molecules ranging in size from small molecules to cyclic peptides [54–58]. These efforts have had various degrees of success, but none has resulted in a useful therapeutic.

1.7 Basis of Nucleic Acid Behavior

Since the importance of generating specific binders for arbitrary DNAs and RNAs should be clear, as should the worth of a binder for the HIV I TAR element, let's look at the challenges associated with the goal of targeting an RNA.

As a general statement, the first step to targeting *anything* is understanding how the target behaves. To understand an RNA target, let's consider what a typical DNA or RNA macromolecule is built from and the general properties of a DNA or RNA macromolecule. while using the former to understand the latter. Finally, we will examine how these emergent properties inform the challenge of creating a binder for an arbitrary polynucleic acid vs. accomplishing the same for a protein of similar size.

1.7.1 Nucleotides

The nucleotide is the monomeric building block of an extended DNA or RNA molecule. A fundamental difference in engineering a protein binder for a nucleic acid vs. another protein emerges from the differences between their monomers: nucleotides and amino acids respectively. Proteinaceous amino acids account for 20 different side-chains with fairly diverse chemical functionality including carboxylic acids, primary amines, amides, guanidino, thiol, thioether, as well as various degrees of hydrophobicity. There are, however, only five canonical nucle-

obases, and only four can occupy a given position on a DNA or RNA polymer. In addition to the limited *number* of possibilities at each position, these nucleobases are limited in chemical diversity (and are of similar hydrophobicity). All nucleobases can be reductively described as small, minimally modified heterocycles, with either purine or pyrimidine as the foundational heterocycle. Figure 1.9 shows the various nucleobases.

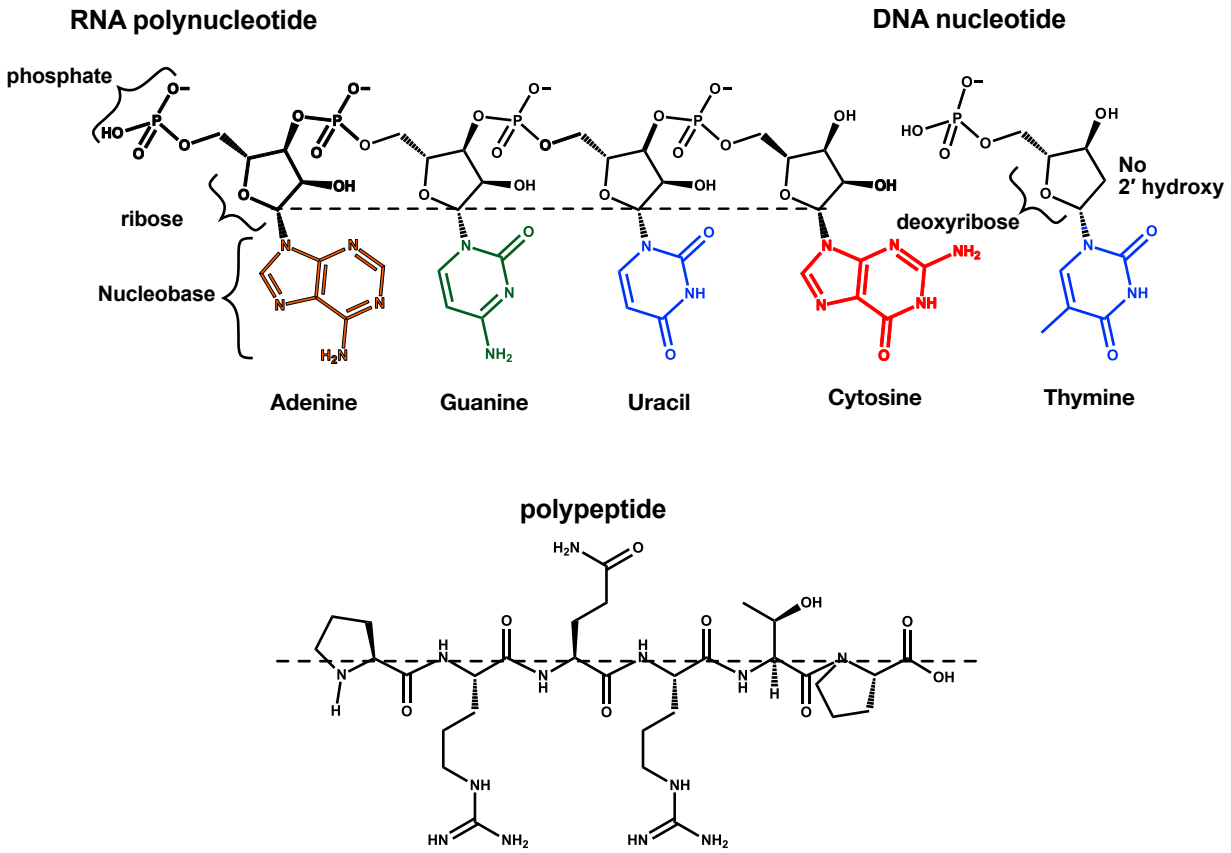


Figure 1.9: Polynucleotide (RNA) vs. DNA vs. Polypeptide) vs. Polypeptide A polynucleotide is made up of nucleotides while a polypeptide is made up of amino acids. The key differences between the two classes of nucleotide are the substitution of thymine for uracil as a nucleobase in DNA and the lack of a hydroxy group on the 2′ carbon in the backbone sugar of DNA (deoxyribose vs. ribose). The primary implications of a nucleotide monomer vs. an amino acid monomer are the minimal chemical diversity of the nucleotides vs. the proteinogenic amino acids, and the largely exposed backbone of a polynucleic acid vs. the internalized backbone of a polypeptide.

These nucleobases are connected to a cyclic pentose, with the glycosidic bond attached to carbon 1 on the pentose. RNA (ribonucleic acid) nucleotides have ribose as the backbone pentose, while DNA (deoxyribonucleic acid) has deoxyribose, which lacks a hydroxy group on the 2′

carbon. This small difference leads to drastically different chemical properties, and affects the way that RNA and DNA interact with the cellular environment and pathways.

The pentose is attached to a phosphate group through the 5′ oxygen. Since these backbone elements are identical on any given DNA or RNA they cannot be used to target a specific DNA or RNA. The nucleobase, sugar, and phosphate together are referred to as a nucleotide (or deoxynucleotide) monophosphate in this form. Importantly, at physiological conditions, the phosphate groups carries a negative charge in both the monomer form and as part of a polymer. This fact is critically important to the emergent properties of a polynucleic acid chain.

Since the sugar and phosphate groups are identical scaffolding connecting the different DNA or RNA bases, they are usually referred together as the sugar/phosphate backbone. Worth noting is that nucleotide monomers with different numbers of phosphate groups (especially adenine mono- di- and triphosphate) are also important signaling molecules, energy sources for reaction catalysis, and enzyme co-factors within a cell in addition to their role as building blocks of complicated polynucleic acids. Therefore, anything which targets an individual nucleotide too specifically will wreak havoc with almost *all* cellular processes.

1.7.2 Nucleic acid as a Polymer String

Though the nucleotide monomers are important, DNA and RNA functionality is predicated upon the polymerization of these nucleotides. This polymerization occurs strictly on the sugar/phosphate backbone, with a phosphate forming a bridge between the 5′ -OH group on one nucleotide and the 3′ -OH group on the next, in what is referred to as a phosphodiester linkage. Nucleic acid sequences are conventionally written in the 5′→3′ direction, since this is both the direction in which they are synthesized by a polymerase, and read by the ribosome [7].

Importantly, if one draws a line through one end of the glycosidic bond, the variable nucleobases will all be on the one side of the line, and the sugar/phosphate backbone will be on the other (as can be seen in Figure 1.9). In contrast, if one draws a line down the peptide backbone of a protein the functional side chains of the amino acids will be found on either side of the

backbone. Thus the non-variable backbone of a polynucleic acid cordons off the variable bases, while in a polypeptide the variable side chains cordon off the non-variable backbone.

This is a gross oversimplification of the actual three-dimensional, structured, reality of both macromolecule classes, but it is a valuable approximation to build on as we move to the next model.

1.7.3 Nucleic Acid as Ribbon

Given the nature of a polynucleic acid as segregated anionic and hydrophobic surfaces, a single-stranded polynucleic acid can be roughly understood as a two-surfaced chemical ribbon with the same essential properties from end to end, but with drastically different properties on either side (illustrated roughly in Figure 1.10).

One side of this two-sided ribbon is the sugar phosphate backbone (Figure 1.10A); the salient characteristic of this side of the ribbon are the regular and identical phosphate groups which give it a uniform negative charge. Due to this negative charge, a polynucleic acid will have general affinity for cations and cationic moieties (notably the side-chains of Lys and Arg residues in proteins) will have fairly strong *non-specific* interactions with a DNA or RNA chain.

The other side of the ribbon can be approximated as a regular series of aromatic loops with the π-systems perpendicular to each other. The variations in this pattern are minor, since any chemical diversity emerges from a limited set of chemically similar nucleobases. The salient characteristics of this side of the ribbon is its hydrophobic character (and therefore its tendency to be buried while in aqueous solution), and the ability of external aromatic π-systems (such as those found on the side-chains of Tyr, Trp, and Phe) to participate in π-π interactions with the nucleotides on this chain Figure 1.10B.

Both sides of the ribbon, of course, participate in hydrogen bonding interactions, but the backbone side has repeating, identical units of sugar hydroxy groups, while the nucleobase surface has heteroatoms on the purine and pyrmidine rings which participate in hydrogen bonding interaction. It is only on this side of the ribbon that there is any variability in the hydrogen

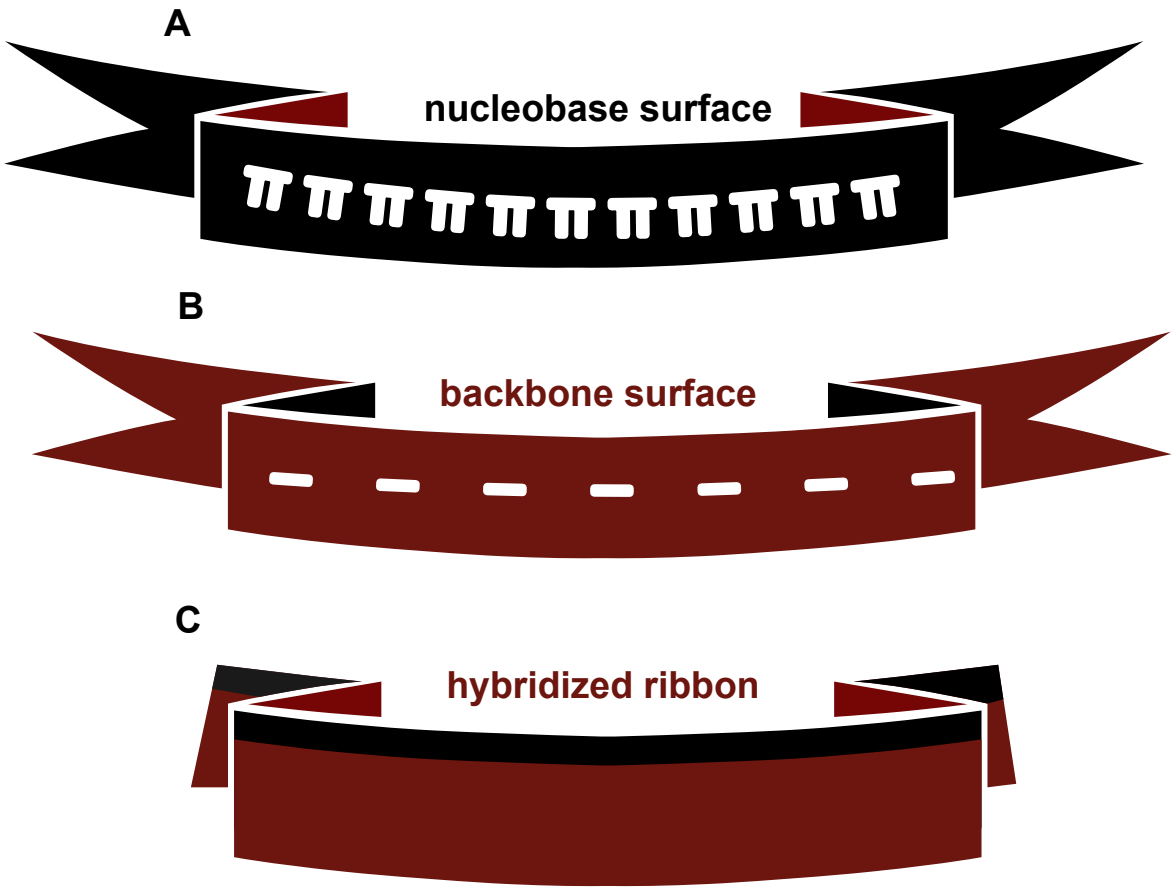


Figure 1.10: Polynucleotide as a Two-Sided Chemical Ribbon Considering a polynucleic acid as a two-sided ribbon is an enlightening mental model. The key point of this model is that each surface has different overwhelming tendencies. The nucleobase surface tends toward interaction with hydrophobic areas, while the backbone side tends toward interaction with cations. A secondary, related, point is that hybridization tends to bury the nucleobase surface, meaning that the majority of exposed surface in a double-stranded DNA or RNA is the charged, hydrophobic, *constant* backbone side of the ribbon.

bonding pattern. Indeed, minor variations in hydrogen bond and space-filling pattern are the *only* chemical difference between polynucleic acids of different sequence.

Considering a nucleic acid in this manner further clarifies the challenge of targeting a polynucleic acid vs. targeting a protein of similar size. When a nucleic acid hybridizes in this model, it does so by two nucleobase surfaces coming together, therefore the nucleobase surface (which is, after all, where *specific* recognition usually occurs) will expose relatively less surface, and the more exposed anionic backbone forms a wall around it.

In contrast, while the peptide backbone is obviously the wellspring of many properties of proteins, it does need not usually need to be considered when targeting a specific protein to the degree that the negatively charged backbone of a nucleic acid does. Any specific binder of an RNA walks a tightrope of anionic interaction. Such a binder *needs* to either interact with or tolerate the drastic chemical property of a persistent and constant negative charge, but it can't interact too strongly with this charge or it will bind *all* nucleic acids rather than a *specific* nucleic acid.

1.7.4 Nucleic Acid Base-Pairing

This "ribbon" model of a nucleic acid also informs an essential emergent chemical property of nucleic acids which is the basis for nearly all of the three-dimensional structure of these molecules—the antiparallel binding of two strands of nucleic acid with complementary sequences. On a conceptual level, this can be approximated by thinking of the hydrophobic sides of two nucleic "ribbons" coming together, burying the hydrophobic bases and exposing the hydrophilic sugar-phosphate backbone.

In the nucleotide level view of a double-stranded chain, each nucleobase participates in specific hydrogen-bonding interactions with a partner nucleobase with the opposite heterocycle foundation, this is known as "Watson-Crick" base pairing (the two hydrogen-bond mediated A·T pair, and the three hydrogen-bond mediated G·C pair). There are also alternative base pairings, such as the "wobble" and "Hoogsteen" base-pairings. These are less favorable than standard

Watson-Crick base, but can still stabilize the structure of polynucleic acids. The different types of base pairing are illustrated in Figure 1.11.

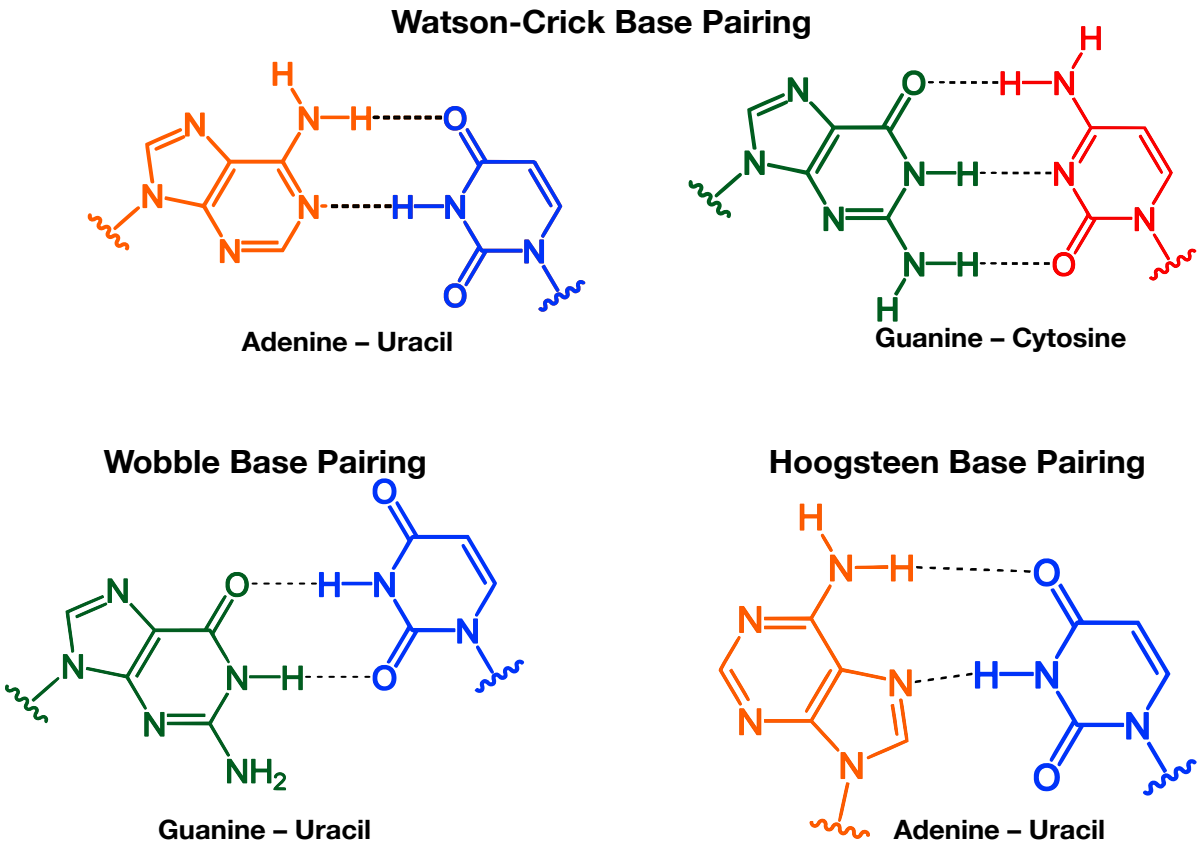


Figure 1.11: Examples of the primary modes of nucleic acid base pairing The most important base pairing interactions are the Watson-Crick A·T and G·C, but Wobble base pairs can pair non-canonical partners, and Hoogsteen pairs can stabilize alternate conformations (as can be seen in Figure 4.4

The phenomenon of hybridization is often thought of as being *driven* by these hydrogen bond interactions, but it is more accurate to say that it is *enabled* by them. This becomes clear when considering the Gibb's free energy equation ($\Delta G = \Delta H - T\Delta S$). The spatially specific binding of two massive biopolymer ribbons is most assuredly entropically costly, but so are the ordered, clathrate-like water formations around the nucleobases of a single-stranded nucleic acid. The hydrogen bonds between purine and pyrimidine bases allow this ordered water to disperse into disorder, making the ΔS value of complex formation less negative than it would otherwise be.

The driving force for hybridization—the ΔH—comes from the hydrophobic effect. When a nucleic acid is base-paired, it minimizes the surface area exposure of the hydrophobic nucleobase side of the ribbon, making the charged backbone the site of the primary interactions. Though it is, importantly, a more predictable and specific structure formation than protein-folding, it is driven by the same forces [59].

1.8 Nucleic Acid 3D-Structures

This tendency of a polynucleic acid to base-pair to a complementary strand leads to the formation of complex three-dimensional structures. Unlike the specifics of protein folding—which remain difficult to predict from a primary sequence—the structure formed by a polynucleic acid stems from specific and predictable interactions between bases. It is here that the properties and tendencies of DNA and RNA (at least in a biological context) begin to diverge.

1.8.1 DNA Structure

For DNA, discussion of biologically relevant structure begins with the familiar double-helix. The double-helix of popular imagination is known as B-DNA. B-DNA is a right-handed double helix, with constant, uneven spacing between the two helices, creating a major groove (22 Å in width) and a minor groove (12 Å in width). The helix undergoes a full rotation every ~10.5 base-pairs. The spatial relationship between bases is repeated constantly with minimal variation [7].

Realistically, this is also where a discussion of the biologically relevant structural variation of DNA *ends*. Single-stranded DNA (ssDNA) is certainly a *chemical* possibility, but is seldom seen in a cellular context outside of the context of DNA replication. Additionally, other forms of double-stranded DNA *exist*, sometimes even in a cellular context, but the proportion of these other forms can be generously described as "trace" [60]. Ultimately, this lack of structural variation makes targeting biologically relevant DNA a simpler, more systematic exercise than targeting complex and amorphous structured RNAs [61]. Though simpler than targeting RNA, the considerations

involved in targeting DNA inform the challenge of targeting structured RNA, and as such one such systematic solution will be described at length in Section 1.13 on page 36.

1.8.2 RNA Structure

RNA, in contrast, demonstrates a more expansive palette of structural variation within the cell. Double-stranded RNA is important to a cell, but unless a cell is infected with a Baltimore Class III viruses with a dsRNA genome, extensive swaths of dsRNA are not of particular concern. Longer strands of RNA in a cellular environment are all-but entirely single-stranded mRNA or essential and conserved tRNAs or rRNA (Table 1.1). And while three-dimensional structural elements are vital in the functional RNAs (such as miRNA or viral RNAs) discussed in Section 1.5, these structured regions are non-extensive enough that they are best thought of as *elements*. Practically, this means that even though a given structured RNA may be *locally* similar to a B-DNA double helix, these areas are generally small enough to bend and breathe in a way that chromosomal DNA does not, and structured RNA molecules cannot be assumed to have the same absolute spatial regularity as a B-DNA helix.

The basic types of RNA structural elements are discussed here and illustrated in Figure 1.12.

Stem-Loops

Stem loops are the fundamental secondary structural element in an RNA. A stem-loop occurs when a series of RNA bases is able to form Watson-Crick base-pairs with a complementary sequence of RNA on the same strand, which is nearby but not contiguous. The paired bases form the stem, while the bases between form a loop. Generally speaking, the loop is between 4–8 bases, though this is sometimes as many as 10 bases (as in U1hpII, an important RNA for this work [62]). A loop at the end of a strand of paired RNA is also called a hairpin loop. A related structural element is the "multi-branch loop," which occurs when stem loops are separated by unpaired RNA strands.

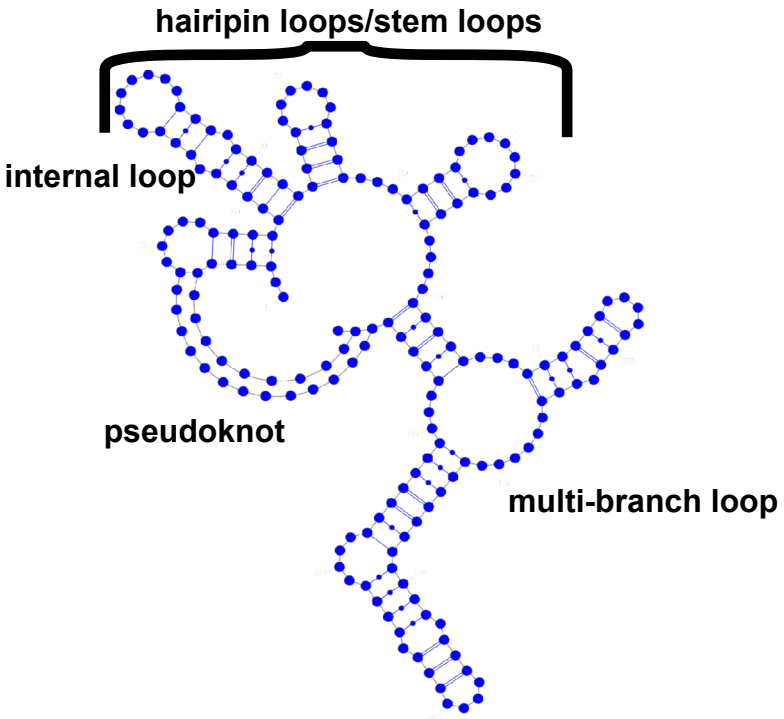


Figure 1.12: RNA Secondary Structure Overview Watson-Crick base-pairing results in a variety of loop-based secondary structure, examples of which are shown here. Important to note is that the base-paired stem regions are largely helical, while the structure elements represent a departure from helicity. Adapted from [63]

Bulges and Internal Loops

1–3 mismatched base pairs on *one* side of a stem form a bulge, which has the primary effect of twisting and changing the direction of the stem. This feature is a hallmark of TAR RNA (Section 1.6). Internal loops are similar, but have concurrent mismatches on *both* strands, which has the effect of distorting the double-helix *without* altering the direction of its axis. Internal loops are commonly associated with miRNAs (Section 1.5.2).

Pseudoknot

Pseudoknots are a sort of secondary/tertiary structure hybrid that occurs when two complementary strands of RNA come near to each other spatially, while being discontinuous (in primary sequence) to other nearby base-paired RNA. This is not a smooth, continuous base pairing, but instead similar to a break on the groove of a vinyl record causing the needle to change grooves abruptly.

Higher Order Structural Elements

Nucleic acids form a handful of specialized higher-order structure. One example is the G-Quadraplex, which is formed by a combination of Hoogsteen pairing of a guanine tetrad. These tetrads can stack via π-π interactions to form G-Quadraplexes. These structures are most commonly found in DNA toward the end of chromosomes as part of the telomere [64], but they do occur in RNA as well. The most notable example of a functional quadraplex occurs in the "Spinach" class of fluorescent RNAs, which utilizes a G-Quadraplex in their small-molecule binding region [65].

Another example of a higher order structure in RNA is a helical stacking (or coaxial stacking) interface. These interfaces are most common in tRNAs and self-splicing introns, and form when two helices wrap around each other. The mechanism is most similar to the splicing of woven rope. These interactions are stabilized by π-stacking interactions between the bases on each helix.

1.8.3 RNA Tertiary Structure

RNA *does* have meaningful spatial tertiary structure, but generally speaking, an extended structured RNA (like the HIV-1 genome Figure 1.8) is well-defined by the structural elements within. This is to say that adding or removing a stem-loop in one position is unlikely to have an effect on a separate stem-loop. Even when the three-dimensional shape is important, it is well-approximated and understood from examining the base-pairing interactions discussed already in this section.

To review the general effects of the common structural elements: generally speaking, the stem of a stem loop structure is a double-helix with a major and minor groove—similar to a B-DNA double helix. Unlike B-DNA, however, mismatches of 1–3 bases are common. These mismatches, as already discussed, warp the helix and change its direction. Mismatches with equal numbers of bases on either side (the internal loops shown in Figure 1.12) tend to warp the helix without drastically altering the direction of the screw-axis through the center, while bulges tend to alter the direction of this central axis. Both distort the general structural features of RNA, and provide a unique shape to each structured RNA. It is important these distortions are important for function (as with ribozymes or riboswitches) [23, 66], or recognition (as with regular bulges on miRNAs) [31].

This variability in structure is the fundamental difference in targeting biologically relevant DNA vs RNA. A profile and top view of dsDNA and various RNAs (including TAR) can be seen in Figure 1.13. Notable are the differences in the types of distortions due to mismatch (i.e. stem loops vs. bulges).

1.8.4 Structure Conclusion

Ultimately, the various structural elements of RNA are somewhat numerous, but they are ultimately more *predictable and finite* than their protein-based brethren. This makes the challenge of targeting them both more and less difficult. The added difficulty derives from the lack of diversity, which makes discerning between target and non-target more difficult. However, this

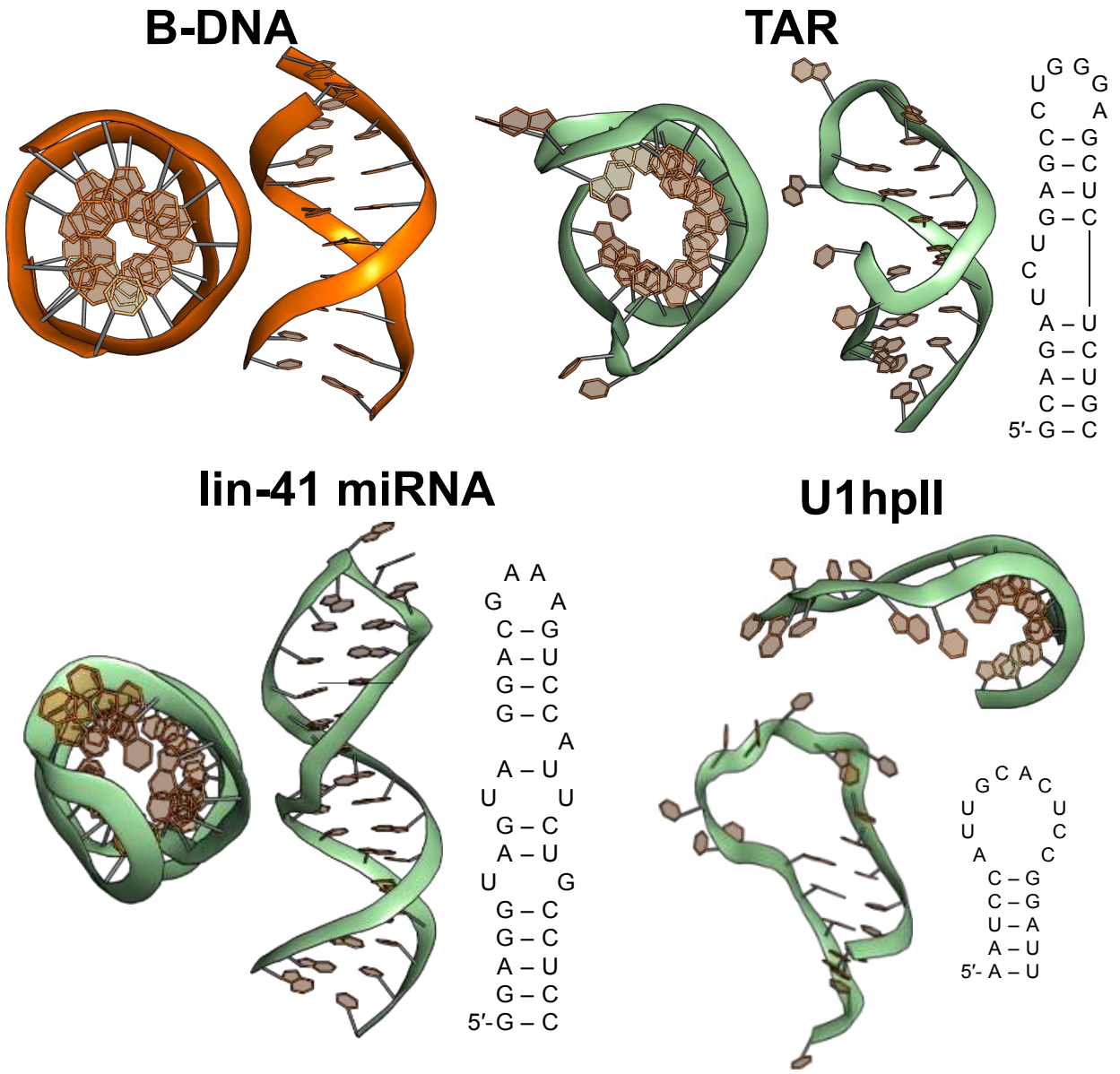


Figure 1.13: DNA vs. RNA Tertiary Structure Comparison A top down and profile view of dsDNA (PDB: 3OMJ), TAR (PDB: 6cmn), lin-41 miRNA (PDB: 2JXV), and U1hpII (PDB: 1URN), with secondary structure diagrams for the RNAs. This demonstrates the consequences of the propensity of RNA to have mismatches in base-paired structures. While biologically relevant DNAs tend to be extremely regular helices, biologically relevant RNAs do not.

feature of RNA also makes targeting an RNA *less* difficult in other ways, due to the increased predictability, and modularity of RNA structure. Once a solution to binding a general class (such as dsDNA and ssRNA) of nucleic acid has been discovered, and the dictates of that interaction understood, they tend too be more effective and adaptable in binding targets across that class (see Section 1.12 and Section 1.14 for examples). Hopefully, this modularity will extend in some form to structured RNAs.

Ultimately, although they *do* have some globularity, extended RNAs are usually considered sequences of structural elements—like beads on a string—rather than an inseparable structural whole, as a folded protein is. A demonstration of this idea can be seen in a representation of the HIV-1 genome, which is traditionally shown as a series of isolated elements, shown in Figure 1.8A.

1.9 Nucleic Acid as Sequence

Finally, an important point that is irrelevant when discussing the chemical abstract properties of nucleic acid, but *sprang* from the elucidation of the structure of DNA: the most important emergent property of polynucleic acids is frequently *information*.

The primary challenge implied by this fact is the difficulty of separating structure and function. When trying to affect a protein, high-order structure and function are intrinsically tied; an ATP binding protein has an ATP binding pocket, or it wouldn't *be* an ATP binding protein. But the DNA which codes for the oncogene *Myc* and the DNA coding for anti-oncogene p53 have identical structures. Likewise, the mRNAs coding for them are identical from a chemical/structural point of view aside from minor variations in spatial positioning of hydrogen bond donors and acceptors. Despite their *chemical* similarity, these two mRNAs lead to opposite effects in a cell, and this divergence ultimately emerges from the information the mRNAs contain, rather than any intrinsic feature of the mRNAs themselves.

1.10 Dictates of Binding Nucleic Acids

Disease relevant RNAs are either a minuscule proportion of the mRNA-ome or one of the "funtional" RNAs shown in Table 1.1, and both of these classes represent a minuscule portion of the total RNA, which is dominated by rRNA and tRNA. Any "therapeutic" that non-selectively affects the wrong mRNA or rRNA is likely to be toxic, rather than therapeutic.

Ultimately, any specific, high-affinity RNA binder must accomplish three primary goals:

- First: an RNA-binding molecule needs to needs to embrace the broad chemical characteristics of an RNA strand in order to generate affinity. It is hard to imagine an RNA-binder without cationic moieties to interact with the negative charge of the RNA backbone, and/or aromatic moieties to participate in π-stacking interactions with an RNA base.

- Second: it must either accommodate the structure of the RNA molecule, or guide the RNA molecule into such an accommodating conformation.

- Third: It must make *specific* interactions to generate selectivity. In practice, this means forming spatially defined, base-dependent interactions (generally hydrogen-bonding, but sometimes steric). Binding specificity based solely on molecule shape, without contact with a nucleobase, seems unlikely.

1.10.1 Small Molecules

Traditional small-molecules (<500 Daltons) have the advantages of cell-permeability, but engineering selectivity in such a molecule is a struggle. This is even more true with nucleic acids than it is with proteins due to the extremely limited chemical variation of nucleic acids, the greater dynamism in the structures, and the general lack of hyrophobic binding pockets.

Certain small molecules *do* bind nucleic acids well. Peter Dervan's lab has designed a small, modular scaffold for binding DNA with pyrrole-imidazole polyamides (see Figure 1.14 and 1.13.6 for further discussion) [67]. The problem of binding dynamic RNA structural *ensembles* has been addressed computationally by Hashim Al-Hashimi's lab [54,68–70], and some random guy in Ben

Miller's lab was able to combinatorially synthesize and screen a library of small molecules to find a binder to the HIV-1 frameshift element [71] using a high-throughput library of aminoglycosides.

The polyamide scaffold from the Dervan lab is shown in Figure 1.14, and generally works by fitting into the minor groove of DNA. Such a solution is not likely to be applicable to a structured RNA, since this binding mechanism relies on the fact that B-DNA has such low structural variation. A small molecule can be built to bind *one* structure to bind DNA, while a binder for an RNA needs to accommodate (and stabilize) *multiple* structures.

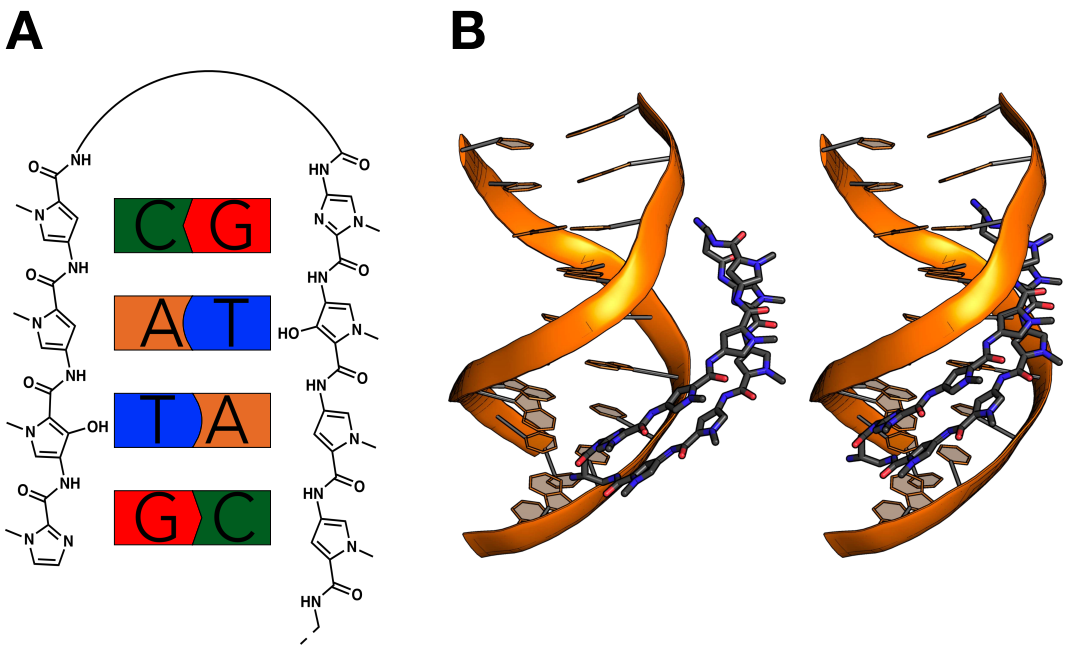


Figure 1.14: Polyamide Scaffold for Binding DNA The general code for the modular dsDNA binding polyamides developed in the Dervan Lab is shown in **A**. Pyrrole/imidazole for C·G pairs, pyrrole/hydroxypyrrole for A·T pairs, hydroxypyrrole/pyrrole for T·A pairs, and imidazole/pyrrole for C·G pairs. **B** shows that that these polyamides bind dsDNA by *fitting* into the minor groove of double-stranded B-DNA.

1.10.2 Nucleic Acids

Sequence complementary nucleic acid is obviously good for targeting RNAs and reducing their expression (see Section 1.3), but this strategy is not suitable for all cases. For instance, catalytic siRNA behaving will eliminate translation, while it may be more desirable to *reduce* trans-

lation. Furthermore, nucleic acids are more prone to degradation in cellular, serological, and therapeutic settings.

1.10.3 Proteins

Proteins are the most natural class of molecules to utilize to accomplish the goals outlined in Section 1.10, since they have a varied chemical and structural toolkit which is naturally compatible for interfacing with nucleic acids. Proteins have the ability to form specific shapes in a variety of sizes, and can bind all scales of structural elements (such as loops, faces, and pockets). This follows, since if we assume the veracity of the RNA World Hypothesis (Section 1.4.1), the fundamental reason they *exist* as a class of molecules is due to their ability to associate with RNA.

Furthermore, proteins exist on approximately the same size scale as DNAs or RNAs on a polymer *and* monomer scale. Therefore the combinatorial modules on the protein (residues) are able to interact with the nucleic acid combinatorial modules at approximately a 1:1 ratio. Proteins are also exquisitely functionally modular—proteins with multiple different functions can be combined into a single genetic unit—for instance the combination of DNA recognition and editing in the TALEN domains discussed in Section 1.13. Proteins also have greater potential to *attenuate* RNA-related events [72], rather than act in a binary fashion, as an siRNA might do.

Proteins are fairly easy to synthesize recombinantly in *E. coli*, and more importantly, protein function (such as RNA binding) is generally easy to screen/select for in an extremely high-throughput manner, as we do in Chapter 2. An important pre-requisite and enabler of this high-throughput screening is the broad availability of existing natural RNA binding proteins which can be used as scaffolds for binding *specific* RNAs.

1.11 Thesis Goal

1.11.1 Develop a Protein-Based Binder for TAR RNA

The goal of the work outlined in this thesis was two-fold. The first, concrete goal was to develop a protein which bound the TAR element from HIV-1 with good affinity and selectivity. The creation of such a protein is inherently useful as a research tool, and may also be a potential source of pharmaceutical therapeutics.

1.11.2 Advance Understanding of Protein–RNA Binding

The second, more abstract, goal was to advance the knowledge of the dictates interactions between RNA-binding proteins and the RNAs they bind. Determining success in this goal will ultimately be determined by the passage of time, but the hope was that we would leave information that can inform and inspire others to new goals and possibilities, just as the current body of functional and structural knowledge informed the work in this thesis.

Before moving ahead to my own research, I'd like take a look back at some case-studies which informed it, in which modular binders for entire *classes* of nucleic acid binding proteins were developed.

1.12 Development of a Modular Binder for dsDNA

1.12.1 Modular DNA Binding Proteins

Given its permanence and clear importance, DNA has been studied and targeted for many years. The benefits of being able to selectively target and effect DNA are obvious—if a gene is defective or detrimental, being able to alter the sequence which codes for that gene will allow a normal and healthy life for the cell. Additionally, if a problem *can* be solved by targeting DNA, then doing so is a fundamentally easier problem, for the simple reason that any DNA edit only has to occur a single time on each DNA sequence in a cell to result in permanent change.

As such, a variety of strategies for targeting DNA have emerged which are instructive for the problem of targeting RNA. All have emerged from the repurposing of well-characterized natural proteins.

Zinc Fingers

One example of a natural class of RNA binding protein is the Zinc Finger domain (ZF). ZFs are a broad class of nucleic acid binding proteins found in Nature, and are somewhat engineerable. They are modular, but the best known classes bind triplets rather than nucleotides. They have been used functionally [73, 74], but the triplet requirement does makes them harder to utilize, since there are 64 possible DNA triplets, and not all have an associated zinc finger [75]. They have been largely superseded by the more modular TAL domains and CRISPR/Cas9 [20, 76].

Leucine Zipper

Leucine zippers (more specifically the helix-loop-helix motif on a single member of a leucine zipper dimer) are usable functionally, but are only somewhat specific in their original forms. This class of proteins has a pattern of basic residues which participate in non-specific interaction at the phosphate backbone interface, and more variable helical regions interact with the major groove/Hoogsteen edge of the nucleobases. Similar to the zinc fingers, leucine zippers are engineerable, but are not sufficiently amenable to engineering to have become widely utilized [75].

1.13 TALENs

Likely the most well-used genome editing construction prior to the advent of CRISPR/Cas9 was the Transcription Activator-Like Effector Nuclease (TALEN). A TALEN is able to specifically target regions of double-stranded DNA for editing by utilizing two proteins fused into a single unit. One part is a general DNA cleavage domain—a non-specific DNA nuclease—and the other is a DNA binding protein which binds specific sequences. It is this second domain,

the Transcription Activator-Like Effector (TALE or TAL), whose tale is most applicable to the problem at hand.

1.13.1 TAL Domain Natural Origin

The TAL domain is derived from *Xanthamonas* bacteria. *Xanthamonas* bacteria infect plants, and have evolved the ability to activate plant genes by contacting the promoter regions for these genes. *Xanthamonas* use TAL proteins, which activate *specific* plant genes that assist the bacterial infection of the organism, and *only* those genes. It was this specificity that intrigued plant biologists, and upon recognized a nuclear localization sequence in the functioning proteins, determined that these proteins were likely operating within the nucleus, presumably upon specific DNA sequences.

1.13.2 TAL Informatics

Even without knowledge of the secondary or tertiary structure of TAL proteins, the primary sequence indicates the modular nature of TAL bindng. Each TAL domain has 17.5 nearly identical repeats (typically 34 residues) corresponding neatly to the 18 DNA base length of the class of promoters they bind. Additionally, the only region on each repeat which demonstrates significant variation within or across TAL genes are positions 12 and 13. These two positions together are known as the Repeat Variable Diresidue (RVD) or, synonymously, the Highly Variable Region (HVR). In 2009, two research groups [77,78] performed a comprehensive survey of TAL effectors and their target promoters and discovered a simple correlation between the identity of the two amino acids in the HVR of a given repeat, and the identity of the DNA base at that position in the promoter. Figure 1.15 shows this simple code. In contrast to Pumilio Repeat Domains (discussed in Section 1.14), all four base-pair possibilities are represented naturally, and a recognition code for 5-methyl Cytosine would later be discovered [79].

The implications of the discovery of this simple code were far-reaching. Rather than having to target an entire *sequence* as a whole, a TAL domain is engineerable by simply choosing a module to match a base, and repeating as needed for the length of the sequence. The discovery

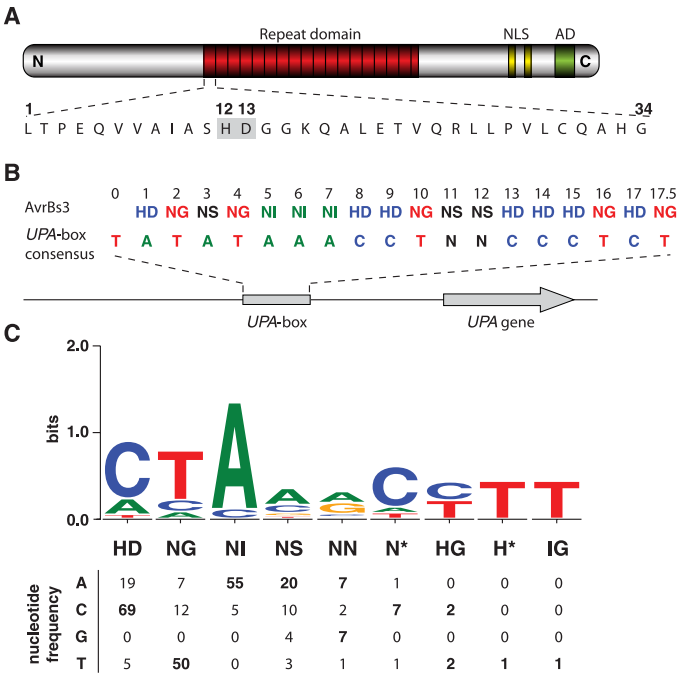


Figure 1.15: TAL Recognition Code Fitting the sequence of TAL domains to the sequence of the promoters they bind revealed that there was a modular code behind this binding. * corresponds to a wild-card residue, and can be any amino acid (adapted from [77]

of this code allowed relatively straightforward modular engineering of proteins to bind to any arbitrary, specific sequence of dsDNA.

1.13.3 Structure of TAL Domain

For proteins, unlike nucleic acids, form and function are largely inseparable. Learning about one will almost always inform and require knowledge of the other. A TAL protein's overall shape is a large, right-handed helical superstructure which traces the major groove of the bound DNA. This superstructure ensures that each repeat, and specifically each HVR pair, is positioned against the corresponding base in the sequence.

Each TAL domain begins with an N-terminal region consisting of a modified version of the canonical repeats, similar in both sequence and structure but containing significantly more positive charge, as well as a tryptophan residue (analogous to the HVR residues in the variable repeats) which participates in a packing interaction with a conserved Thymine base present at the beginning of all TAL binding regions. This combination of positive charge and hydrophobic

packing generates a certain amount of non-specific affinity which is built upon, with specificity, by the following repeats.

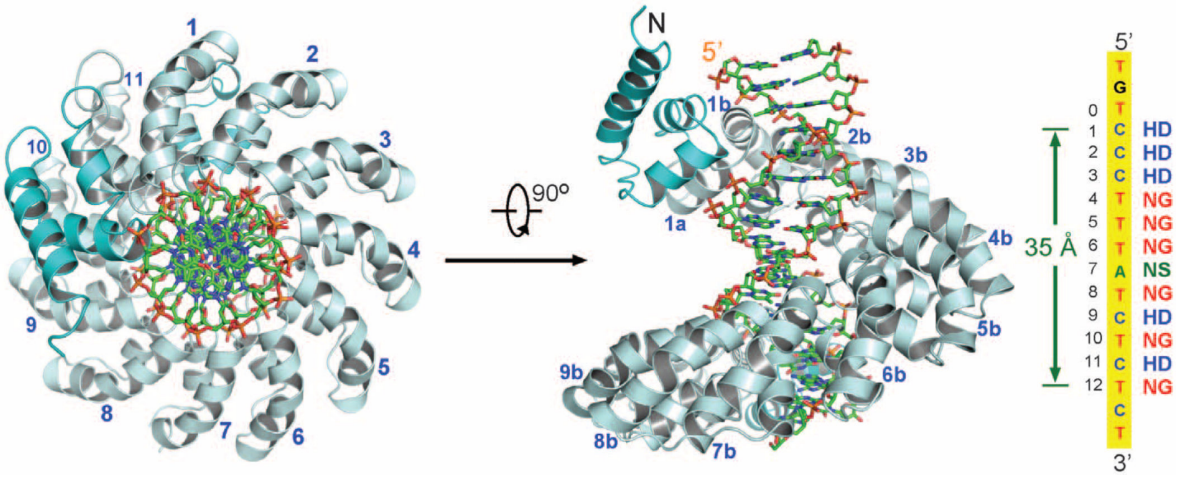


Figure 1.16: Structure of a TAL domain (dHax3) bound to dsDNA The superhelical structure of a TAL domain binds to the major groove, while the RVDs make specific contacts. The RVD/DNA translation is shown o on the right. This TAL domain contains 11.5 repeats with flanking N- and C-terminal helices shown in cyan. Two perpendicular views are presented, with the DNA duplex shown in sticks. Adapted from [80]

The structure of these repeats has evolved to perform the single purpose of aligning the HVR residues into position to contact their respective bases. Each repeat consists of two left-handed helices connected by a loop that contains the HVR residues. The bundled helices fit roughly into the major groove of a B-DNA helix, which positions the HVR domains in contact with the paired bases perpendicular to their Watson-Crick interactions. The HVRs themselves consist of one structural residue (generally N for an "A" or "T" contact or H for a "C" or "G" contact) that lines up the second residue. The second residue participates in either a base-specific contact (based on hydrogen bonding or steric interactions), an exclusionary "interaction" which results in specificity by having three of four base possibilities be untenable, or a "wild-card" interaction in which the HVR can accommodate *any* base.

1.13.4 TAL Domain Engineering

The major triumph of TAL domain engineering was in deducing, from the function, sequence, and structural features of TAL proteins, that they could represent a modular "code" for binding arbitrary DNA sequences. The natural code as-is has been quite robust and usable. Within a year of the publication of the initial code, TAL domains had already been engineered into the TALN genome-editing tool. This code has also proved amenable to further optimization for specificity by screening the small library made by the two bases in the HVR.

1.13.5 TAL Domain Lessons

The TAL domain is the most robustly modular of proteins that recognize dsDNA, and the discovery and engineering provide insights which will prove key for the analogous problem of engineering RNA binders.

- First: the more directly usable the solutions developed by Nature over billions of years of evolution, the better. In this instance, a human-usable modularity was possible, even trivial, since Nature's solution was already modular.
- Second: form follows function—the shape of a protein and the shape of the nucleic acid it binds need to rhyme. For a large, broadly featureless molecule like a nucleic acid, spatial specificity is vital.
- Finally: affinity is best achieved through large-scale interacting with features common to *any* nucleic acid strand, while specificity is then derived from smaller-scale interactions. Just as the non-specific hydrophobic effect *drives* nucleic acid hybridization, but precise hydrogen-bonding enables it, so too do protein binders need a driving force to bring about an interaction *and* a sop against entropy to make that interaction last.

1.13.6 Binding DNA vs. Binding RNA

Though there are lessons to be learned from engineering binders to DNA there are also caveats in extending these lessons to RNA binders. These caveats primarily derive from the

fact that DNA has repeatable, extended, and *consistent* structure. The best way to illustrate this paradigm is in the fact that TAL domains are able to bind a *single* DNA conformation (a B-DNA double-helix), and in so doing bind *DNA*. Disease-relevant RNA, by contrast, has multiple conformations, which implies an entropic cost to get to the *correct* conformation for binding.

An extreme example of the simplification that this assumption of consistent structure can provide is demonstrated by the small-molecule polyamide binders developed by Peter Dervan's lab [67] (discussed previously in Section 1.10.1 and shown in Figure 1.14). These minimal small-molecules based on polyamides bind DNA with a simple modular substitution for each Watson-Crick pair, since they can be designed to match the consistent and predictable B-DNA double helix. In essence, a DNA binder can be engineered to *match* a single structure. It is more difficult to build such a binder for unstructured RNAs, or disease-relevant structured RNAs, which do not have *a* structure, but exist as a dynamic ensemble which must be dynamically matched or stabilized.

1.14 ssRNA binding

1.14.1 Introduction to ssRNA

Aside from the low entropic cost, a complementary way to consider the amenability of double-stranded DNA to modularly defined binding proteins is the ease of pattern recognition. The chemical variability of the bases may be minimal, but at least this minimal variation occurs at predictably in space. Therefore, the binding problem can be reduced to matching one of four hydrogen bond and/or steric patterns, then expanding that pattern in a constant spatial interval; this is essentially pseudo base-pairing. For RNA this assumption of regular, exposed, bases is true in one important class: single stranded RNA (ssRNA). Since ssRNA conforms to this assumption of regular spacing between bases, it is amenable to being coaxed into a pseudo base-pairing interaction.

As with TAL domains, the most straightforward way to accomplish these pseudo base pairing interactions is through repurposing of the solutions that Nature has already achieved. The three

most broadly studied classes of ssRNA binders are Pumilio/fem-3-binding factor (PUF) (widely known as Pumilio Repeat Proteins), Pentatricopeptide Repeats, and Zinc Fingers. Of these, PUF proteins are the most important, since they represent an all-but universal solution to the problem of binding ssRNA. We will examine why PUF proteins have proven so amenable to engineering by tracing the route from initial discovery to straightforward tool.

1.15 Pumilio Repeat Proteins

1.15.1 Natural Origins of Pumilio Repeat Proteins

Pumilio was initially known as a factor involved in abdominal patterning in *Drosophila* embryos [81], with an unknown mechanism. The gene for a Pumilio Repeat domain was first sequenced in 1995, and by 1999 it was confidently hypothesized to be involved in transcriptional regulation by binding mRNA [82]. The general fact that these proteins consisted of eight near-identical repeats was clear, but the implication that the repeats may correspond to one-to-one repeat/base binding was not understood until the crystallization of Pumilio proteins from humans and *Drosophila* with a cognate RNA [83–85].

1.15.2 Structure of Pumilio repeats

A Pumilio Repeat protein binding a single stranded RNA takes on the shape of a gentle arch, which can be seen in Figure 1.17A. The RNA meets the underside of this arch with the Watson-Crick edge of its nitrogenous bases, the bases slightly splayed from the curve. The repeated, modular, structure is readily apparent, with each repeat spaced perfectly to facilitate contact with a single RNA base. Canonically, natural PUF proteins consist of 8 nearly perfectly repeated units, with imperfect repeats on the N and C termini. These terminal repeats are more positively charged than the canonical repeats, and likely provide general RNA-binding functionality.

Each of the canonical repeats consists of 36 residues in three α-helices and associated loops. and is identical save for 3 variable positions within a series of five residues contained on a single alpha helix. This region is sometimes denoted as 12XX5, since the 1st, 2nd, and 5th residues vary

between repeats (the Xs represent constant positions on each repeat). Each repeat unit consists of three right-handed α-helices. One helix contacts the RNA at a nearly perpendicular angle to the RNA backbone, while the other two helices and loop regions act as a structural backbone and somewhat cationic shell. Before the structure of a PUF protein/RNA interaction was solved, it was hypothesized (reasonably) that the cationic residues were directly contacting the anionic RNA.

Instead, the actual interaction reveals a striking similarity to that of double stranded nucleic acid. The relatively hydrophobic face of the PUF protein contacts the hydrophobic RNA bases, while the charged residues on the shell face out toward the water. Thus, just as in the interaction between two complementary nucleic acid strands, the interaction between the protein and RNA is driven by the hydrophobic effect. A side-effect of this feature is prominent in PUF proteins with more than the usual 8 repeats. PUF proteins with 16 repeats make a pronounced c-shape when not bound to an RNA as the hydrophobic face minimizes solvent contact, but upon burying the hydrophobic face into an RNA, the bound protein/RNA combination once again forms the gentle arch associated with PUF proteins (see Figure 1.17).

1.15.3 Mechanism of Pumilio Repeat Binding

The binding mechanism in PUF proteins is fairly simple, and our understanding of it has remained largely unchanged since its elucidation. In each repeat, residue 2 in the 12XX5 motif is involved in a non-specific hydrophobic/cationic interaction with the RNA chain. This non-specific interaction involves Arg, His, Tyr, or very rarely Asn residues. Arginine has a bulky side-chain that can interact with the hydrophobic RNA bases, and terminates in a positive charge, which may participate in Hydrogen bonds or draw in the anionic RNA backbone. The less common asparagine has a similar profile of long side-chain terminating with a hydrogen bond participant. His or Tyr can both intercolate their aromatic side-chain into the π-system. Whatever the identity, this residue at position 2 is known as the "stacking residue."

The residues in positions 1 and 5 participate in specific interactions with their base. These interactions are either Hydrogen bonds or steric Van der Waals interaction; for instance, an Asn in position 5 and a Gln in position 1 participate in a total of three hydrogen bonds with a uracil base. These residues are analagous to the HVR residues in a TAL domain, and will be referred to as "binding residues" henceforth. Worth noting is that while not all of the important interactions between binding residues and RNA are hydrogen bonds, the residues are generally capable of forming hydrogen bonds. An example of a PUF/RNA interaction can be seen in Figure 1.17.

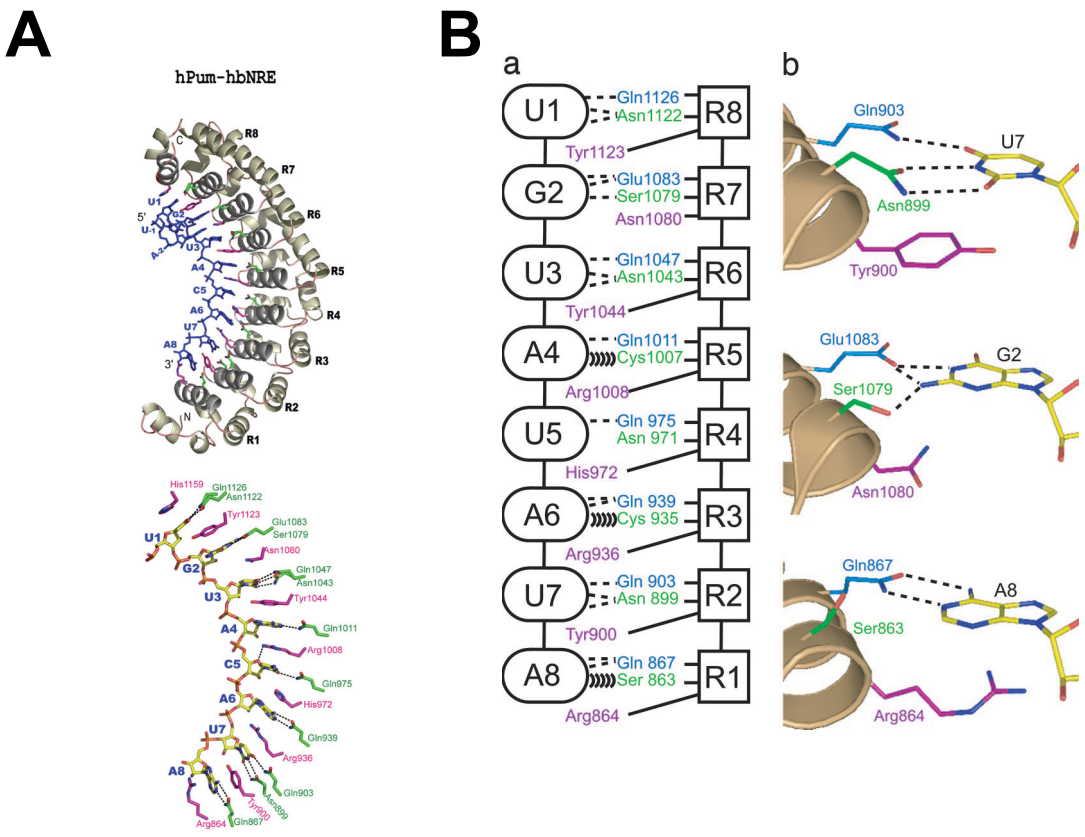


Figure 1.17: Structure of a Pumilio Repeat Domain A the structure of a Pumilio Domain is an arched superstructure made up of α-helices which position the variable residues into place to bind the RNA, and **B** the code for and illustrations of the variable residues which make specific contacts with RNA bases. Dashed lines correspond to Hyrdogen Bonds, parentheses to Van der Waals contacts. Adapted from [86]

1.15.4 Engineering Pumilio Repeats

The first attempt at algorithmically altering the Pumilio scaffold to bind a non-native RNA was a qualified success [86]. The authors showed that altering the PRD HsPUM1-HD protein predictably alters the specificity of the RNA sequence it binds. For instance, replacing the two adenine-specific recognition residues in repeat 3 (Cys and Gln) with the two uracil-specific recognition residues from repeat 2 (Ser and Glu) resulted in a protein which binds a G2U RNA with ~55-fold better K_D vs. the wild-type RNA. This same strategy resulted in similar changes in specificity for G→U, U→G, and A→G point mutant RNAs. This strategy also yielded the important insight that since the aromatic/cationic contact residues of were interacting with *general* properties of RNA, using Tyr, His, Arg, or Asn was to some degree irrelevant to the overall RNA binding properties. Certain stacking residues would prove to be *better* than others, but none had a binary effect on RNA binding.

The qualifications in this success were that the engineered proteins had excellent specificity for their new cognate RNAs, but still demonstrated a worrisome variation in affinity (~100-fold difference between the A→G and G→U protein/RNA pairs). But the most important issue to address was, of course, the lack of an algorithmic repeat which could bind Cytosine.

1.15.5 Cytosine Binder Evolution

The development of a cytosine-binding PUF protein is a textbook example of the targeted use of high-throughput screening to complement rational design. Though high-throughput screens are by definition random, these screens must be carefully designed if they are to yield good results.

The fundamental considerations in designing a high-throughput screen are how much randomness is desirable and whether to confine that randomness to a specific region. If the randomness is to be combined, *where* it should be confined is of critical importance. In the case of the highly-modular Pumilio repeat domain, the strategies already used to algorithmically engineer the protein directly suggest how to design a high-throughput screen. The bases on an RNA

8-mer interact with Pumilio repeats in a spatially defined one-to-one manner independent of position (that is, the "code" for binding is based on target residue *identity*, rather than placement). Since the wild-type PUF protein used as the primary basis for engineering natively binds the RNA sequence 5′-uguauaua-3′, a variant which binds 5′-ugCauaua-3′ would only be different on the sixth repeat, which means only ~30 residues are now considered. To narrow it further, only 3 of the ~30 residues on the sixth repeat form contacts with the RNA, and only 2 of those residues are expected to participate in *specific* interactions. This is the beauty of modularity in protein engineering—knowledge of the model and mechanism of binding allows a massive combinatorial problem to be narrowed down to 2 bases, accounting for a manageable $20^2 = 400$ possible proteins.

To analyze these variants in a high-throughput manner the authors used a variant of the well-established Yeast-2-Hybrid (Y2H) assay, which operates by expressing the target of binding and the possible binder as genetic fusions. In this case, the target and binder are a DNA-binding and a DNA-activation domain respectively. These fusions are then produced, using cellular machinery, in the cytoplasm of a yeast cell. If and only if the probed interaction occurs do the DNA-binding and activation domains come together and enable expression of a colorimetric reporter gene, a gene necessary for survival of the yeast, or both. In this case, a successful binding event triggers both a histidine synthesis gene which allows the yeast cell to grow on a deficient plate, and the commonly used colorimetric LacZ reporter gene.

The challenge of using a yeast hybrid screening method to identify a protein–RNA interaction is that it is impossible generate a yeast-expressible protein–RNA fusion. This is addressed using a known, sequence specific, protein–RNA binding interaction (derived from MS2 bacteriophage) as a link [87]. In this case, the LexA DNA binding domain (which binds a specific DNA sequence) was fused to the MS2 bacteriophage coat protein (which binds a specific structured RNA, here called MS2 RNA). To perform the screen, MS2 RNA was fused to the Cytosine-containing target RNA (5′-ugCauaua-3′). The LexA-MS2 fusion binds the MS2-target RNA fusion, thus forming a bridge between the activation domain and the target RNA. The PUF protein library was

fused to a DNA *activation* domain. Therefore, only if the PUF variant binds the target sequence (containing Cytosine) will both the DNA-binding and DNA-activation domain be in place to lead to transcription of a gene required for histidine synthesis, as well as the *LacZ* gene. The system is illustrated in Figure 1.18. This use of an RNA adapter in the Y2H assay is commonly known as a Yeast-3-hybrid (Y3H) assay [88]. The Pumilio Repeat Domain, with two randomized residues corresponding to binding of position 3, was expressed as a fusion with the Gal-4 Activation domain. Yeast which grew on His-deficient plates were analyzed, and striking sequence homology was found [89, 90].

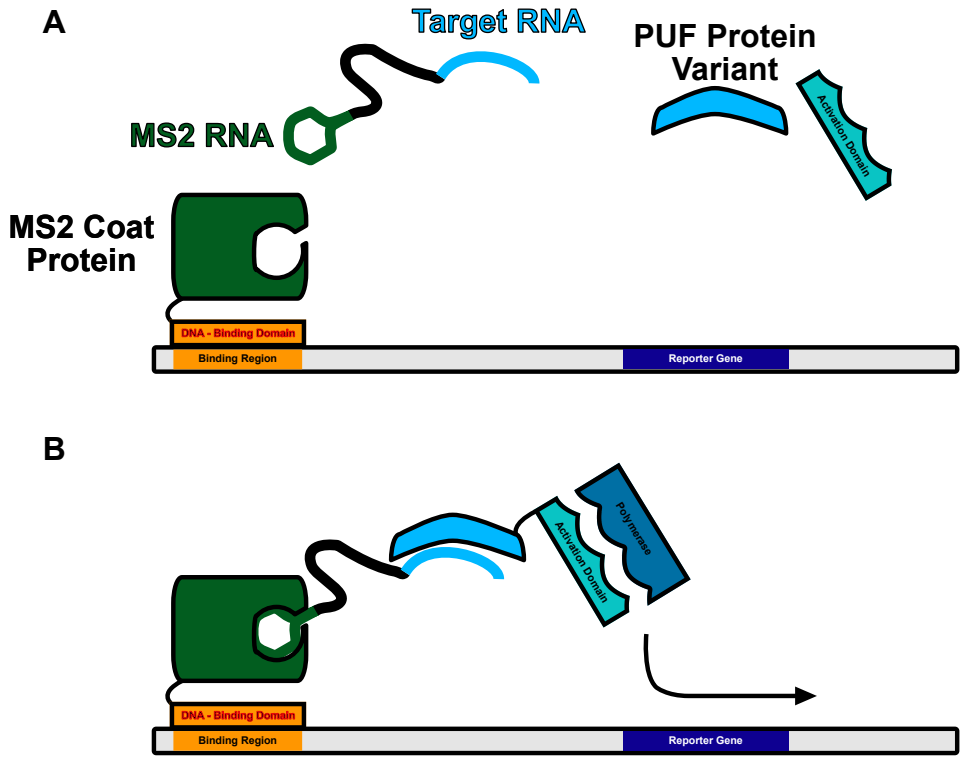


Figure 1.18: Yeast-3-Hybrid for Finding Cytosine-Binding Pumilio Repeat The **A** components and **B** assembled Yeast-3-Hybrid system used to screen for, and find, a cytosine binding Pumilio Repeat. Adapted from [88] using the components in [89].

The final determination was that the best specific binding residues for C-binding Pumilio repeat are Ser and Arg, with Tyr used as the non-specific contact residue. This motif can be grafted on to any of the repeats in a PUF protein, modularly enabling cytosine binding. Interest-

ingly, this determination informed the discovery of possible natural PUF proteins with cytosine binding repeats [89].

This illustrates the ability of high-throughput screening to fill in the gaps left by Nature. The fundamental idea of using *targeted* randomness to build on what Nature has already evolved, forms the basis of affinity maturation work in Chapter 2.

1.15.6 Pumilio Repeat Domain Utilization Arc

The utilization of Pumilio Repeat Domains traces an interesting arc. First, observation and engineering lead to a simple idea and dream: a protein scaffold that could be used to generate a binder to an RNA as easily as designing complementary RNA strand. Knowledge and observation led to a potential method, and experimentation led to a viable proof of concept. This proof of concept was elaborated on *because* it is already viable, and increasing amounts of study and engineering led to more complexity than might be wished. Eventually the knowledge and understanding reached a critical mass which enabled the full realization of the simplicity of the original dream. The "pumby" domain discussed in this section represents the modular PUF protein developed to the point of ultimate simplicity.

The initial efforts at reprogramming Pumilio Repeat Domains were only somewhat modular, given the large variation in binding affinity between different reprogrammed PUF proteins [86]. This variation in binding affinity occurs since each repeat on PumHD (the PUF protein engineering basis) and its close analogues binds its base in a slightly different manner. As such the eventual "code" derived solely from the natural proteins was imperfect, and involved one of four different different possibilities at each repeat. Figure 1.19 shows this code.

Golden Gate assembly (chosen due to its use in studying TAL repeat domains, since similar engineering challenges require similar tools) was used to rapidly screen variants of each Pumilio repeat, and a simple, modular code (shown in Figure 1.19) was discovered [91]. A single binding residue pair was determined for each nucleotide, and Tyr was discovered to be a universal non-specific stacking residue. Interestingly, later work focused on structural analysis would de-

termine that substituting Tyr for the uncommon His and Asn stacking residues improved affinity in all examples studied [92].

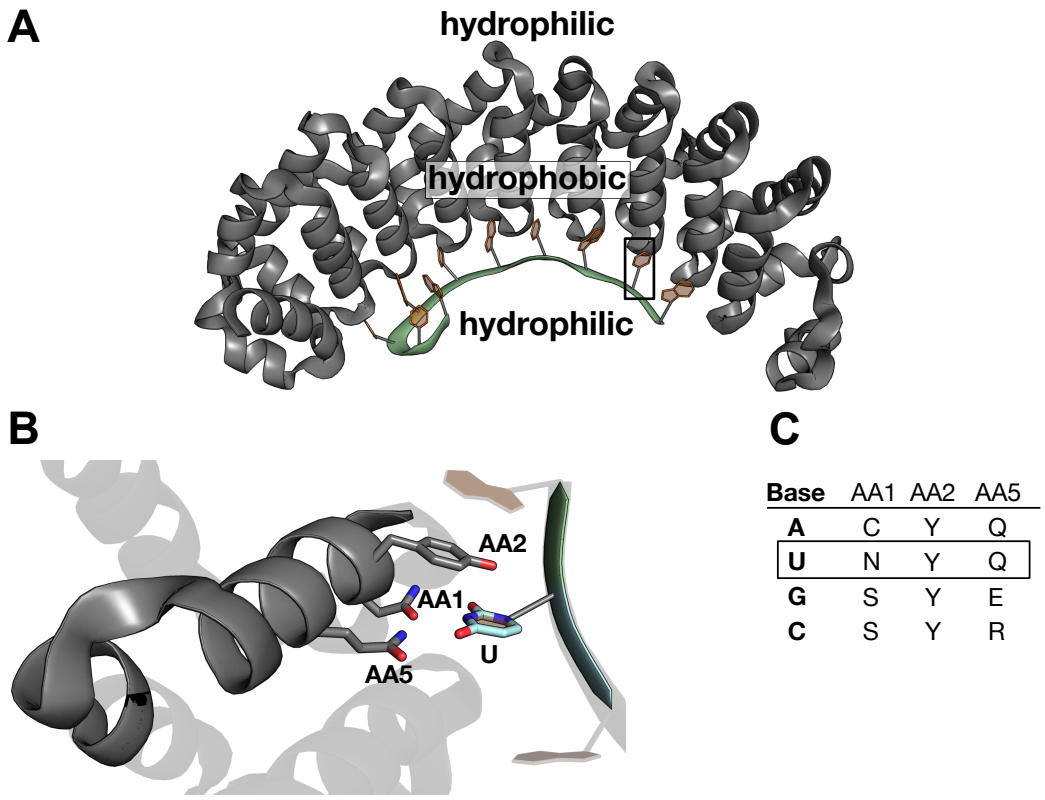


Figure 1.19: Modular "Pumby" Domain Code A An overview of a PUF protein with hydrophilic and hydrophobic faces marked. **B** A detail of a simplified "pumby" domain, here shown with a Uracil binding repeat, is the ultimate expression of modular PUF protein engineering. AA1 and AA5 participate in base-specific interaction, while AA2—the "stacking" residue is *always* Tyr in the pumby system. **C** shows the full code for the modular Pumby system from [91], PDB: 1M8Y

The modularity of PUF proteins is not dependent on the strict, canonical model of 8 Pumilio repeat domains binding an 8-mer RNA. For example, a PUF protein with 16 repeats binds a 16-mer RNA with approximately the same affinity as a PUF protein with 8 repeats binds an 8-mer RNA, and exhibits similar tolerance for mismatched bases. This plateau does indicate a limit to affinity, but does not change the essential advantage of a reasonably broad size range available for engineering binders to specific sequences of varying lengths.

1.15.7 Other ssRNA binders

Pentatricopeptides

There are other binders of ssRNA that operate in a modular fashion. The most notable of the alternatives to PUF proteins are the Pentatrico peptide Repeat proteins (PPRs), which consist of repeated dual-helical units ~30 residues in length which form a superhelix. PPRs have some advantages over PUF proteins: they are more numerous in the genome of eukaryotes, meaning that there are more possible foundations for building an arbitrary binder. Most importantly, the initial rough code found naturally has a solution for binding to Cytosine, something which PUF proteins initially lacked. PPRs are also more apparently scalable, given there are natural PPRs with up to 30 repeats, presumably binding RNA chains 30 bases in length.

However, these advantages don't outweigh the problems. PPRs have a discernible recognition "code," but it is less clear than that of PUF proteins. The tolerance for mismatches at any given position is unpredictable, and the contribution of each repeat/RNA base interaction varies. The repeats near the 5′ end of the RNA are more important to binding than those near the 3′ [93]. Probably most importantly, Pentatricopeptides are difficult to study closely due to the lack of a known crystal structure. This means that the excellent understanding of structural vs. contact residues of PUF proteins isn't accessible in pentatricopeptides.

Zinc Fingers

Zinc Fingers (ZFs) have been engineered to bind dsDNA, but of the most familiar "CCHH" type ZFs some in the TFIIIA family do bind ssRNA.

Ultimately, RNA-binding zinc fingers have the same issues as their DNA-binding brethren, namely that they tend to bind short *motifs*, rather than exhibiting truly complete modularity [75, 94, 95].

1.16 Structured RNA Binding Proteins

1.16.1 Background

The development of a modular code for PUF proteins to bind ssRNAs is indeed a great achievement. Combined with the lessons learned from TAL domain engineering, this development provided a simple, though daunting, set of consideration for the eventual design of binders to arbitrary *structured* RNA.

- First: Modularity is key to usability and development
- Second: Structural and functional knowledge are necessary to discern and inform that modularity

Pre-existing natural PUF proteins left the problem of modular RNA recognition all-but solved due to the fact that they are *naturally* and *inherently* modular. The primary sequences alone provided the hint and there were numerous good crystal structures which allowed mapping repeat to base. Once these requirements were in place, the engineering required to build a tool that could be understood and adapted by an RNA biologist was time-consuming, but relatively straightforward.

It is unlikely that there is a single class of protein that can bind *any* arbitrary structured RNA, but a "flow-chart" style modularity may be possible. A hypothetical possibility of what form this might take follows: general attributes such as secondary structure *type*, such as "stem-loop" could be matched to a class of RNA binding proteins. Once the initial structure type has been matched, somewhat more specific attributes like loop size, or bulge locations in the stem, can be matched to a particular scaffold within that class, and sequence could be matched with individual H-bonding or packing residues.

As has been noted, the variety of RNA secondary and tertiary structure is daunting, but it is *finite*. Furthermore, unlike the seemingly infinite considerations that determine the shape proteins, the mechanisms by which RNA *folds* into its various structural possibilities are well understood. One possibility implied by this reality is that longer, more extensive RNAs (like the HIV-1

5′-UTR) may be targetable by modular combinations of smaller proteins which match the individual elements within the extended RNA (proposed in an excellent review by Lunde, Moore, and Varani [61]). If this is true, the challenge of RNA binding would be all but solved as long as the modular combinations for the (again, daunting but finite) possibilities of structure and sequence for every folding variety of RNA ~25 in length were known.

Given this paradigm, the first step to finding a good binder for structured RNA is to pick a scaffold based on existing binders to structured RNAs. Fortunately, many natural RNA binding proteins exist to provide insight. Examples of these structured RNA binding proteins include the Muscle-Blind (MBNL) proteins which bind extended CUG repeats, NOVA domains which bind stem-loops, and PIWI and Staufen domains which bind double stranded RNA. Examples of these proteins can be seen in Figure 1.20.

1.16.2 RNA Recognition Motifs (RRMs)

By far he most common, and most studied class of RNA binding proteins is the RNA Recognition Motifs (RRMs), at least one of which is included in ~2% of the proteins coded by the human genome [97–99]. They have a compact structure which folds well, and are defined by their βαββαβ structural elements, with the four β-domains making up a compact structural face. The RRMs have some general chemical features as well: they are highly positively charged, which provides some necessary general affinity, and there are conserved aromatic residues on the β-face that participate in non-specific π-stacking interactions with RNA bases [99].

Unlike the PUF domain, which binds ssRNA exclusively, or something like the Staufen domain shown in Figure 1.20, which binds dsRNA exclusivley, RRMs *(*primarily) bind ssRNAs, but sometimes bind dsRNAs. The RRMs as a class are structurally well-studied, with over 30 crystal structures in the PDB. A few of these RRM/RNA interactions are shown in Figure 1.21. This abundance of proteins, and the concomitant abundance of crystal structures means that there are many potential starting points. RRMs usually have some baseline non-specific affinity of

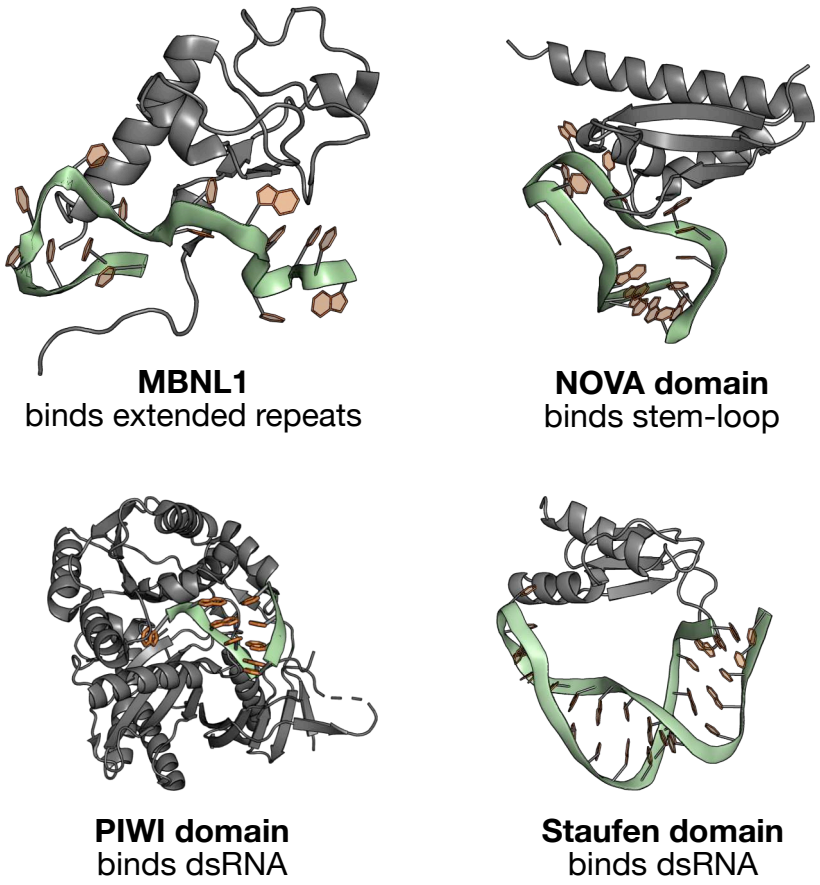


Figure 1.20: Example Classes of RNA Binding Proteins Shown in this figure are varieties of RNA binding proteins. Though these classes are not used in this work, it is worth contemplating the variety of RNA forms which have existing solutions that may be viable to build upon. For more information see [96]

RRMs for RNA means that, for better and for worse, a certain amount of interaction can be expected for any given RNA.

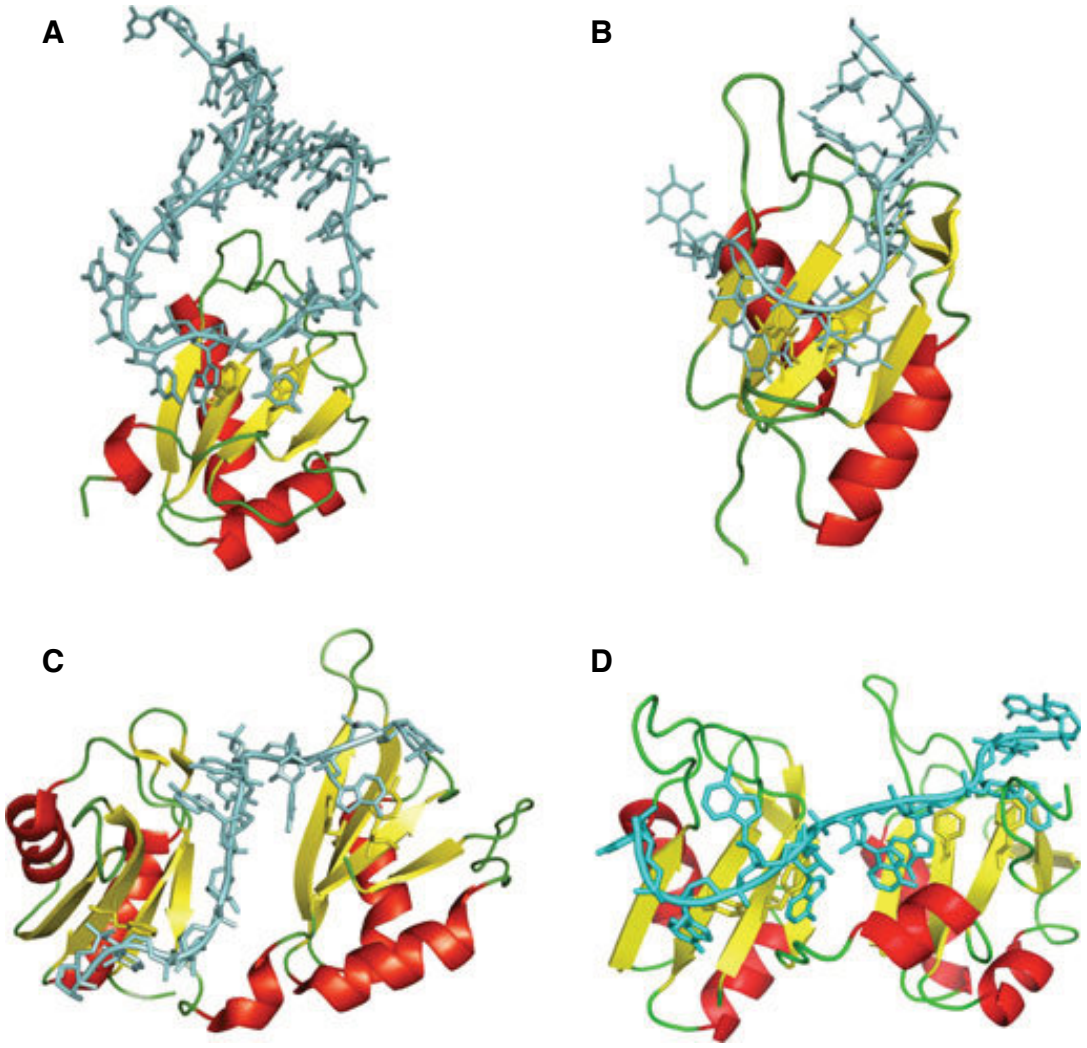


Figure 1.21: Examples RNA Recognition Motif Proteins Structures of RNA-recognition motifs (RRMs) in complex with their RNAs. The proteins are shown as ribbons, (red: helices; yellow: sheets; green: loops). Conserved aromatic residues on protein binding surfaces are highlighted. **A** Single RRM recognition: crystal structure of the U1A complexed with U1hpII (PDB: 1URN) and **B** Solution structure of the RBD of Fox1 in complex with UGCAUGU (PDB: 2ERR) and **C** Crystal structure of an AU-rich element recognized by the HuD protein (PDB entry: 1G2E) and **D** Crystal structure of the poly(A)-binding protein in complex with polyadenylate RNA (PDB: 1CVJ). Adapted from [61]

Finally, the benefit of all this knowledge is that any given interaction is well understood. The downside, so far, has been that these binding interactions are too complex to be engineered well. A 2013 review article from the Varani Lab summarizes the problem well, and is quoted here:

"So far, no RNA recognition code has been defined for RRM domains, and no successful attempt has been made to engineer RNA-binding sequence specificity using RRM. For RRMs that bind RNA specifically and with high affinity, the extensive and non-contiguous RRM–RNA interfaces present a challenge in constructing a library of mutant RRMs that is large enough to cover all the protein residues that potentially contribute to RNA binding. In fact, when the structures of U1A–RNA complexes were first available, it was found that the RNA–protein interface more closely resembled protein–protein recognition than DNA–protein recognition [100]. Furthermore, specific base recognition is often achieved through protein backbone interactions, a feature that limits the range of RNA sequences that may be targeted [61]."

The underlying problem in binding structured RNA is too much complexity with too many dependencies. If one is to engineer such a protein, one should choose a well-understood one, and we chose the best-understood of this well-studied class: the N-terminal RRM on the U1 Binding Protein (U1A). U1A exists on both sides of the structured/unstructured RNA binder divide. Since U1A's recognition sequence is for *unpaired* 5′-AUUGCAC-3′ RNA, it is technically correct to describe it a ssRNA binding protein, but since this sequence is only recognized in the context of a stem-loop, it is *also* accurate to say that U1A recognizes a structured RNA, and is similar to a dsRNA binding protein.

In any case, U1A this unusual binding mode is to an unusual cognate—the U1phII RNA. The recognition sequence is in the middle of an unusually long 10-base Hairpin Loop (though a loop of that size bears no resemblance to a hairpin). And this unusual target is bound with unusual affinity; the K_D of the U1A–U1hpII interaction is ~40 pM, a level of binding associated usually associated with the likes of biotin and streptavidin [101].

1.16.3 U1A–U1hpII Mechanism of Binding

The general mechanisms of binding for this protein are well understood. The first U1A–U1hpII co-crystal structure was solved in 1994 [102], and has been well-studied since. A crystal structure can be seen in Figure 1.21A, with an annotated structure in Figure 2.1. U1A is well stud-

ied enough that we have a good understanding of the nature of the involvment of every piece of the protein.

For instance, it is understood, down to the individual residue, the role played by the many lysines and arginines in a two-step process of general recruitment followed by specific recognition [103,104]. It is also well-known that the Tyr13 and Phe56 (part of a conserved element known as the RNP face (a name derived from ribonuclear proteins where the element was first observed), are vital in complex stability, and that these conserved residues are *general* rather than specific binders of RNA. Also well-characterized is the helical region near the C-terminal. This helix is not a usual part of the canonical RRM fold, and it acts as a sort of mobile clamp which forms a hydrogen bond network on the RNA opposite the protein [103]. Another well-understood region which is the β1α1 loop, which plays a role in specific RNA contacts [105].

The most important structural element of U1A is the β2β3 loop, the role of which is well-understood [106]. The β2β3 loop is unstructured in U1A crystals, and only becomes structured when co-crystallized with the U1hpII cognate RNA. The β2β3 loop extends into the U1hpII loop, making up a disproportionate amount of contact between U1A and U1hpII RNA. Shortening the U1hpII loop abolished binding [106,107], indicating that there was a special relationship between this unusually high degree of contact between the protein loop and the unusually large RNA loop.

Perhaps most importantly, both high-throughput and fully rationalized engineering on the U1A protein had already been accomplished by way of the β2β3 loop. For instance, a high-throughput screen with randomization of β2β3 loop residues resulted in a U1A variant with improved U1hpII affinity [108].

Rationalized engineering predicated upon size discrimination had also been accomplished prior to our work. This engineering was based on the fact that U1A does not bind a mutant of U1hpII missing two bases in its stem-loop. Shortening the β2β3 loop by one base does not restore binding, but compensating for the two missing RNA bases by deleting *two* bases at the end of the β2β3 loop (K50 and M51) restored some degree of affinity [106,107]. Though unspectacular on its

own, this simple experiment demonstrates a basic confirmation of understanding: if the system is perturbed, restoring it is somewhat predictable.

1.16.4 U1A E19S

In the course of trying to engineer novel RNA binders, Brett Blakeley, my eventual mentor, was able to validate these results from the Laird-Offringa lab, and became interested in the concept of the β2β3 loop as a steric ruler. If a region plays such an extended and fundamental role, surely there must be further engineering to be done... Simultaneously, active work was begun to engineer an RRM to bind a *disease relevant* RNA hairpin, with TAR being among the most desirable targets.

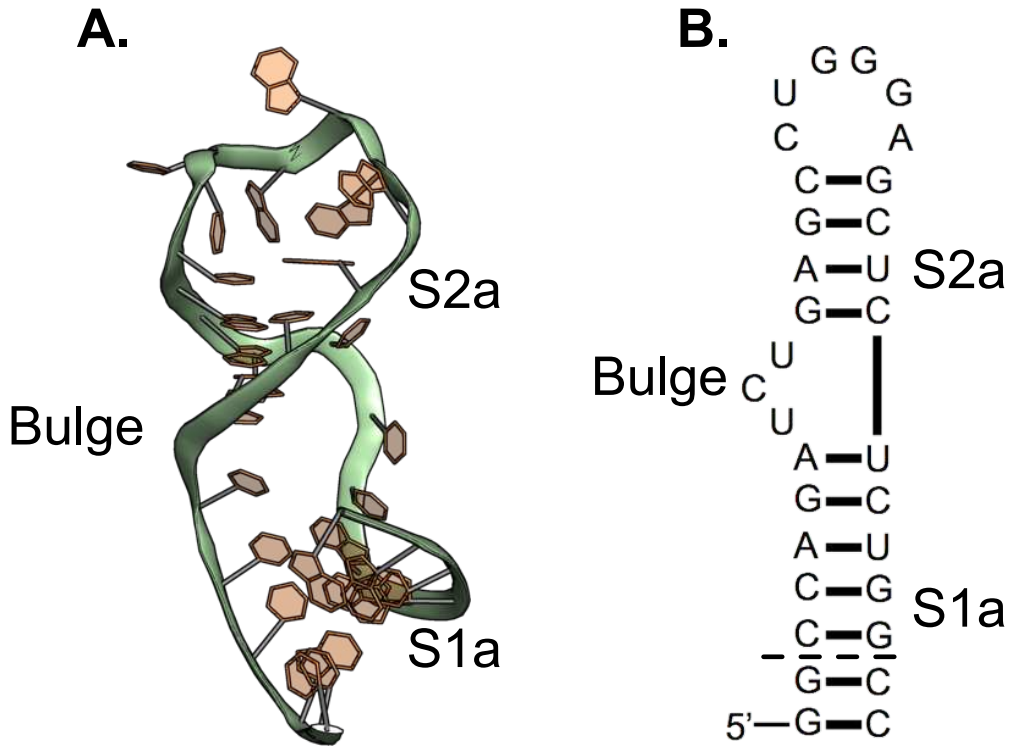


Figure 1.22: The HIV I TAR Element A The crystal structure of the bulged stem-loop of HIV I TAR RNA [45] and **B** The sequence and putative secondary structure of isolated TAR RNA targeted in this work. The lower stem (S1a), upper stem (S1b), and distinctive UCU bulge are labelled. The two G·C pairs below the dotted line are not part of the TAR element, but rather a clamp to ensure that the small RNA maintained its hairpin structure.

To this end, Brett was able to find a single point mutation—changing the glutamic acid on the hydrogen-bonding β1α1 loop to a serine—which greatly lowered the affinity of U1A for U1hpII, and in fact made U1A somewhat selectively bind TAR. [105]. The process of finding this (only slightly) improved binder was quite involved. Brett had had to purify and analyze single mutants. The fact that he found an improved binder at all demonstrates another important necessity for success in research—luck. It was decided that finding a truly excellent binder for TAR was going to require more drastic changes to the protein involving synergistic mutations, which would require high-throughput screening of a large, randomized library (a process known as affinity maturation). I joined the project just as this affinity maturation began.

Chapter 2

Affinity Maturation of U1A E19S for TAR RNA Binding

2.1 Chapter 2 Introduction

2.1.1 Chapter 2 Summary

U1A is the best studied RNA Recognition Motif, with extensive mapping of region to function. Previous work in our lab leveraged this extensive knowledge to generate a point mutant (E19S) with specificity for the TAR RNA element of the 5′ UTR region of the HIV genome. In order to make an avid and selective binder for this RNA, Brett Blakeley and I used yeast display to express and screen libraries based on U1A E19S. Our initial library had 5 randomized positions (46, 48–51) in the β2β3 loop region of the U1A protein (3.2×10^6 possible proteins). After three rounds of selection, the library was diversified by randomizing either the β1α1 region or the C-helix region of the protein. The C-Helix library proved the most promising, and was screened for three more rounds. The sequences were analyzed, and found to have a highly convergent β2β3 loop with the consensus sequence of **P**R**TRTP** (R47, unbolded here, was left unchanged from the original protein). We initially attempted to measure the avidity of our screened library proteins using techniques already established in our lab (Fluorescence Polarization and Yeast Display), and though we were unable to firmly quantify the TAR Binding Protein–TAR interactions, it seemed clear that some of our selected proteins had excellent affinity for TAR.

2.2 Chapter 2 Attribution

This chapter is adapted from [109].[1]

Brett Blakeley, of the McNaughton Lab, was primarily responsible for the library design and screening strategy, with help and input from myself. My role increased roughly chronologically as I took over the project from Brett.

Brett designed the positive control experiments, and observing preparation of same was my introduction to yeast display. The initial library design was Brett's, though I assisted in preparation and added some microbiological advice and assistance. While Brett had final choice in screening conditions, I also provided input on these conditions, and I helped prepare library sorts. I provided input and assistance in the diversification of the library.

Once the library sorting was complete, I learned Fluorescence Polarization from Brett and performed the assays shown in Figure 2.9. I assisted in the yeast display assays shown in Figure 2.10A, and independently performed the assays shown in Figure 2.10B. I was responsible for the sequencing and library analysis shown in Table 2.4.

2.2.1 Chapter 2 Background

The field of Biochemistry revolves around interactions. Life at the cellular level is a massive cascade of logic-gated cause and effect driven by the selective coming-together of large macromolecules (as well as the interactions of these large macromolecules with their partner small-molecules). Both studying and affecting Biochemistry frequently involve engineering synthetic binding molecules in the hopes of observing, disrupting, or encouraging these interactions. As *de novo* protein design remains a daunting challenge, our lab utilizes a strategy of "semi-design" in which we modify existing proteins that already perform the *general* function we desire (in this case, binding to an RNA stem loop), and engineer them to perform a *specific* subset of that general function.

---

[1] Crawford, DW, Blakeley, BD, Chen, PH et al. An Evolved RNA Recognition Motif That Suppresses HIV-1 Tat/TAR-Dependent Transcription. *ACS Chemical Biology*, 11(8):2206–2215, **2016**

Our goal was to develop a specific and selective binder for the TAR element of HIV I, which is vital for the proliferation and infectivity of the HIV virus and as such is a desirable therapeutic target. The specific scaffold we build upon in order to achieve this goal was the best-characterized of the common RNA Recognition Motifs, the "A" subunit of the human U1 ribonucleoprotein (hereafter known as U1A or wtU1A), which is involved in human mRNA splicing, and selectively binds the U1hpII RNA with exquisite affinity (K_D = 40 pM) [102, 106].

Brett Blakeley, in previous work, had shown that a single mutation from Glu → Ser on position 19 on the U1A protein could broaden, and to some degree "switch" affinity. U1A E19S binds U1hpII with a K_D of ~ 12 μM, while binding TAR with a somewhat tighter K_D of ~4 μM [105] Realistically, these low μM K_D values indicate *broader* specificity, i.e. a protein with *general* affinity for stem-loop RNA. The goal of this work was predicated upon the fact that this broadened specificity of U1A E19S made it a good scaffold to use for building a protein with truly *altered* specificity for TAR RNA. Position 19 had been chosen for study because it was known that the β1α1 loop in which it is contained participates in specific hydrogen bonding interactions with U1hpII. The hope was that by using our knowledge of the protein, we could use high-throughput screening to produce a more specific binder of TAR RNA.

The logical starting point is the region which contributes the most surface area to the U1A–U1hpII interaction: the β2β3 loop. It was well-established that the β2β3 loop is disordered in crystal structures of the U1A protein alone, but in co-crystal structures of U1A and U1hpII (PDB: 1urn) that it fits well into the abnormally large 10 base loop of the U1hpII RNA. Furthermore, it was well understood due to work by both our lab and the Laird-Offringa lab, that reducing the size of the β2β3 loop on the protein creates a protein which binds a shorter RNA hairpin loop, indicating that the β2β3 loop acts as a sort of steric ruler. [106, 107].

Also important to this work, it was known that the "C-Helix"—an α-helical region near the C-terminus of the protein outside of the canonical βαββαβ structural elements of an RRM—acts as a clamp to stabilize the protein–RNA interaction [110]. The regions of U1A important in binding U1hpII are highlighted in Figure 2.1A, the co-crystal structure of U1A–U1hpII is shown

in Figure 2.1B, and a comparison of the sequences and secondary structures of TAR and U1hpII is shown in Figure 2.1C.

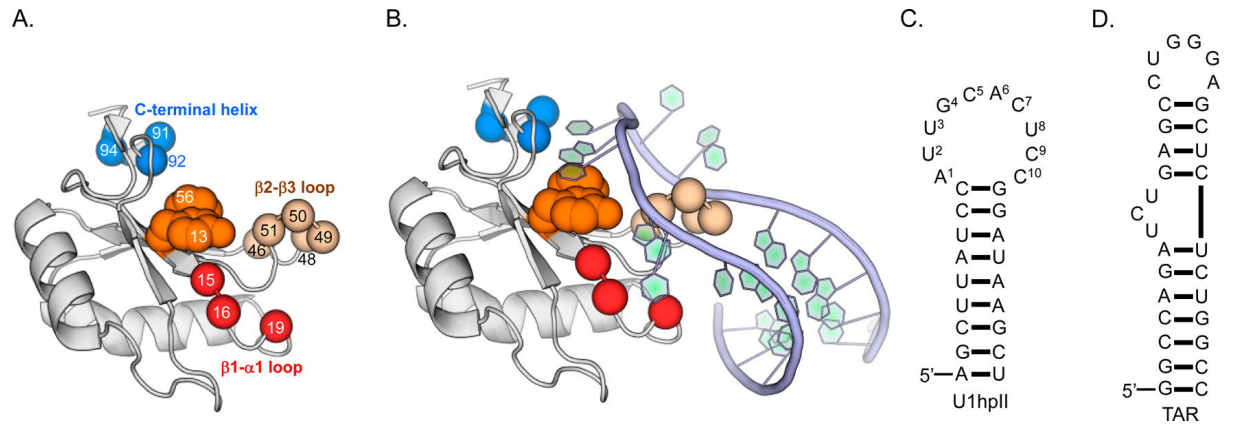


Figure 2.1: Crytal Structure of wild-type U1A binding U1hpII and comparison between U1hpII and TAR A Crystal structure of U1A (PDB: 1URN) highlighting four major binding hot-spots, the RNP face (residues 13 and 56) involved in general stabilization, the β1α1 loop (residues 15, 16 and 19) involved in hydrogen bond contacts with U1hpII, the C-terminal Helix (residues 91, 92, and 94) which acts as a mobile clamp, and the β2β3 loop (residues 46, 48–51) which constitutes the majority of binding **B** U1A bound to its natural U1hpII cognate **C** U1hpII secondary structure compared to **D** TAR Secondary Structure

2.3 Design of Screening Strategy

We suspected that with U1A E19S we had achieved the reasonable best-case scenario for rationalized semi-design; no other point mutant of U1A would be a significantly better starting point. We decided that our next engineering step would be to use a high-throughput screen to find an avid TAR binder. After all, the ability to easily make and screen large libraries is among the greatest strengths of proteins as research tools and pharmaceutical leads. Furthermore, it was already established that U1A was at least somewhat amenable to high-throughput selection [108].

2.3.1 Screening Method

The first consideration was which of the many screening techniques to use, and there are many indeed, ranging from simple microarrays, phage display, mRNA/ribosome display), bacterial display, and yeast display. We rejected the idea of microarray screening since we lacked

the instrumentation and expertise to utilize it. We also rejected the idea of using mRNA display due to our lack of experience with such a specialized technique. Phage display, though previously used for U1A, has several drawbacks, notably the requirement of a solid-state scaffold (sub-optimal for selection), while bacterial display was not considered usable for displaying something as large as U1A. We chose, for a variety of reasons, to move forward using yeast display.

Yeast Display

Yeast display originated in the Wittrup lab, and utilizes the existing Aga1/Aga2 membrane proteins in a yeast cell. The Aga1 and Aga2 proteins are expressed from different transcripts. The Aga1 protein embeds in the membrane, while the Aga2 protein connects to the Aga1 protein through two disulfide bonds. Approximately $5–10 \times 10^4$ Aga2 fusion proteins can be displayed on an individual yeast cell. The protein of interest can be expressed as a fusion to Aga2, and multiple affinity tags can be included. Display can then be confirmed with the use of antibodies specific to these affinity tags. Therefore, the protein of interest is available for interaction with a target, and since the interacting protein is tethered to surface of a cell, the sequence which codes for the protein can be retrieved and amplified simply by allowing the cell to reproduce and later extracting the DNA. Figure 2.2 shows a diagram of this method.

Yeast display has some intrinsic advantages for this particular screen. The first is that the technique itself is operationally simple—*S. cerevisae* yeast take more time to grow than the *E. coli* used in phage display, but they are more robust in both their growth and induced display. Furthermore, genetic manipulation of yeast—including transformation and DNA recovery—is straightforward, making library creation and diversification relatively simple. Yeast, being eukaryotes, have chaperone proteins, which reduces concern about aggregation and mis-folding of proteins. Finally, unlike phage display, yeast display pairs well with Fluorescence Activated Cell Sorting (FACS), which is high-throughput, robust, and was well-known by our lab at the time. Even more importantly, the fluorescent data gives meaningful affinity data *during* the screening

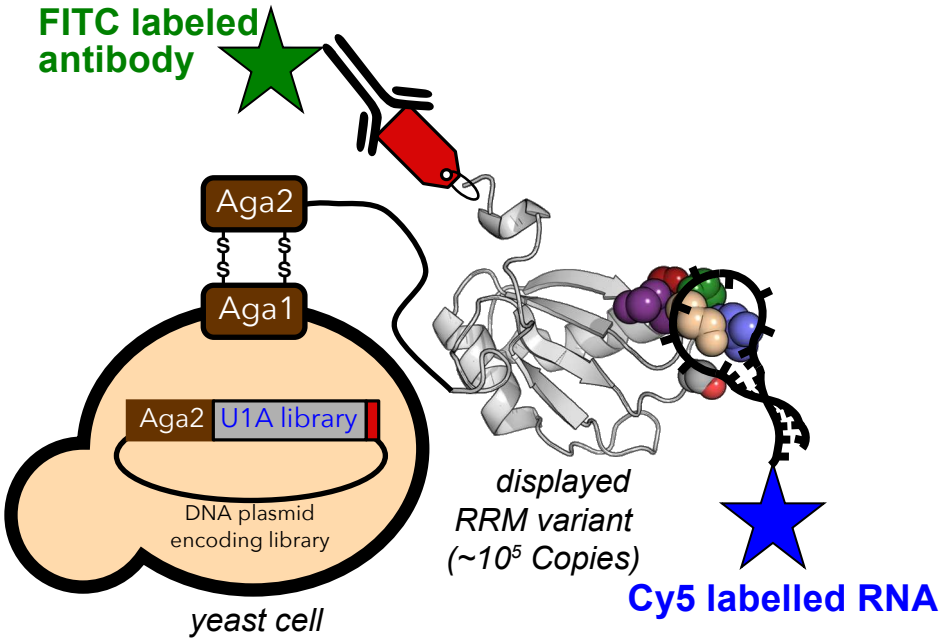


Figure 2.2: Overview of the Yeast Display Technique The yeast display system used to screen our U1A E19S-based library can be seen here. The main features are the ability to screen a U1A based protein with affinity tags to ensure display.

process, allowing more nimble and thoughtful adjusting of screening protocols, as well as reducing the possibility of enrichment of false-positive "cheater" sequences.

2.3.2 Screening strategy

Since the screening technique defines the upper limit of library size, a concurrent step with the selection of screening method is deciding which residues on a protein to randomize. We were planning on screening a library made up of variation in one of the four main binding hot-spots on U1A. Our choices were the β1α1 loop (residues 15, 16, and 19), the β2β3 loop (residues 46–51), the conserved binding face (notably residues 13 and 56), and the C-Helix (residues 91–94). The conserved binding face seemed best kept conserved, our hope was that the β1α1 loop was already somewhat optimized for TAR binding via the E19S mutation, and since the C-Helix place a role in the *second* step of the binding process, it was an illogical place to begin selection. Therefore, we chose to optimize the β2β3 loop.

The upper limit of yeast transformation is around 5×10^7 transformants, and in order to have a reasonable expectation of a complete library, the general rule of thumb is to have ~ten-fold more transformants than possible members of the protein library [111,112]. Mutating all six residues on the β2β3 loop would lead to an unreasonably large library size ($20^6 = 6.4 \times 10^7$), and the arginine at position 47 is known to play an important role in general RNA recruitment [106], similar to residues 13 and 56. Leaving out R47 meant we had a more reasonable 5-member library ($20^5 = 3.2 \times 10^6$) consisting of residues 46 and 48–51.

2.4 Yeast Display Confirmation

Prior to making the yeast library, we first needed to confirm that yeast were indeed capable of displaying U1A variants, and that U1A was still able to bind RNA while displayed on the surface of a yeast cell.

2.4.1 Cloning U1A Variants into Yeast

The components of the yeast display platform are the specially engineered *Sacharomyces cerevisae* variant EBY100 and the pCTcon2 plasmid (which encodes the Aga2 protein, cut sites for cloning a fusion with the *Aga2* gene, an Amp-resistance gene for selection in *E. coli*, and a Trp syntheis gene for selection and maintenance in *S. cerevisae*. Both were generously provided for our use by the Wittrup lab. The yeast are also available via the American Type Culture Collection ATCC (MYA-4941) and the plasmid is available from AddGene (41843).

Wild-type U1A (wtU1A) and U1A E19S genes were amplified via PCR templated by wtU1A and U1A E19S in pET plasmids with FWD NheI U1A (5′-ATA TAG CTA GCA TGG CCC AGG TGC AGC-3′) as a forward primer and REV BamHI U1A (5′-CGG GAT CCT GCG GCC GCA ACC-3′) as a reverse primer. The pCTcon2 Plasmid was digested using NheI and BamHI (High-Fidelity versions from New England Bioloabs) and treated with Calf Intestinal Phosphatase (CIP). The amplicons were digested with BamHI and NheI and purified. The amplicons were then ligated into the digested pCTcon2 vector using a Quick Ligation kit from New England Biolabs, and this

ligation transformed into chemically competent 5α *E. coli*(NEB C2978). The transformation was plated (following 1 hour of rescue in SOC media at 37 °C) onto LB plates containing 100 µg/mL Carbenicillin (obtained from GoldBio).

This construct represents, from 5′ to 3′, a sequence encoding the Aga2 protein, Human influenza hemagglutinin (HA) tag, a GGGGSx2 linker, the U1A variant, and a *myc* affinity tag (Sequence: Section C.1.2 for wtU1A and C.1.3)

Yeast Electroporation

The plasmid DNA was isolated using a standard DNA miniprep kit (from Omega Biotek), and transformed into EBY100 yeast made electrocompetent using a standard yeast electroporation protocol [113], using a GenePulser (Bio-Rad) and Gene Pulser/Micro Pulser Cuvettes 2 mm (BioRad). Electroporator was set to 540 V and 25 µF. Transformed yeast were rescued using YPD media, and plated on SD-CAA plates.

All of the following cloning and yeast transformations in this thesis were performed using the above protocol unless otherwise noted.

2.4.2 Positive Controls

The transformed yeast were cultured in 250 mL baffled shaker flasks containing 50 mL of SD-CAA minimal media (5.4 g/L Na2HPO4, 8.6 g/L NaH2PO4 • H2O, 20 g/L dextrose, 6.7 g/L yeast nitrogen base lacking amino acids, 5 g/L casamino acids, 200 kU/L penicillin, 0.1 g/L streptomycin), which lacks tryptophan, until they had reached mid-log phase, at which point they were centrifuged and resuspended in 50 mL SG-CAA (5.4 g/L Na_2HPO_4, 8.6 g/L NaH_2PO_4 • H_2O, 1 g/L dextrose, 19 g/L D-galactose, 6.7 g/L yeast nitrogen base w/o amino acids, 5 g/L casamino acids, 200 kU/L penicillin, 0.1 g/L streptomycin) media (which induces production of the Aga2 fusion protein) to a cell density of 1.0×10^7 cells/mL [113]. Simultaneously, yeast containing wtU1A DNA were re-suspended to a density of 1.0×10^7 cells/mL in SD-CAA. The assumption was that the uninduced yeast would serve as a proxy for the surface of a yeast cell to ensure that EBY 100 did

not have any affinity for TAR RNA without our displayed RRM. Yeast were induced for 24–36 hours. The above is the standard yeast induction protocol used throughout this thesis.

To analyze display, 10^6 yeast cells were centrifuged, the buffer aspirated, and the pellet washed in a 1.7 mL eppendorf tube with 1 mL of ie-cold Phosphate Buffered Saline (PBS), washed with 1 mL of ice-cold Phosphate Buffered Saline (PBS) containing 1 mg/mL Bovine Serum Albumin (BSA). Cells were resuspended in 1 mL of this PBS-BSA containing a 1:10,000 dilution of FITC conjugated anti-*myc* antibody (Abcam ab117599, which was used extensively throughout this thesis). Additionally, wtU1A samples were incubated with various concentrations of U1hpII RNA labeled on the 5′ end with a Cyanine-5 fluorophore (Cy-5, RNA ordered from IDT). Following 1 hour of incubation rotating at room temperature the cells were pelleted, washed once with PBS-BSA, and left as pellets on ice after removal of the PBS-BSA.

Melt and Refold RNA

As a general note, before any assay involving RNA (yeast display, FP, ELISA, etc.) the RNA was melted and refolded. First, the RNA was brought to a lower concentration to prevent dimer formation by diluting the 100 μM stock into water to a concentration of 10–20 μM depending on the assay. The RNA was heated for 2 minutes by submersing the tube in boiling water (~95 °C). Following this two minutes of heating, the tube was immediately plunged into ice.

Flow Cytometry

After <20 minutes (the time needed to transport samples to flow cytometry), cells were resuspended in 1 mL PBS-BSA, and analyzed for fluorescein fluorescence via Flow Cytometry (Beckman-Coulter MoFlo). Results for both antibody binding and RNA binding for wtU1A and U1A E19S can be seen in Figure 2.3.

Both wtU1A and U1A E19S seemed to be displayed on ~20% of yeast cells, which was in line with normal display, and the wtU1A seemed to bind U1hpII RNA. The uninduced yeast cell surface bound neither the antibody nor the Cy-5 labelled U1hpII RNA, which was the confirmation

needed to finalize our decision to use yeast display, and we proceeded with generating our library.

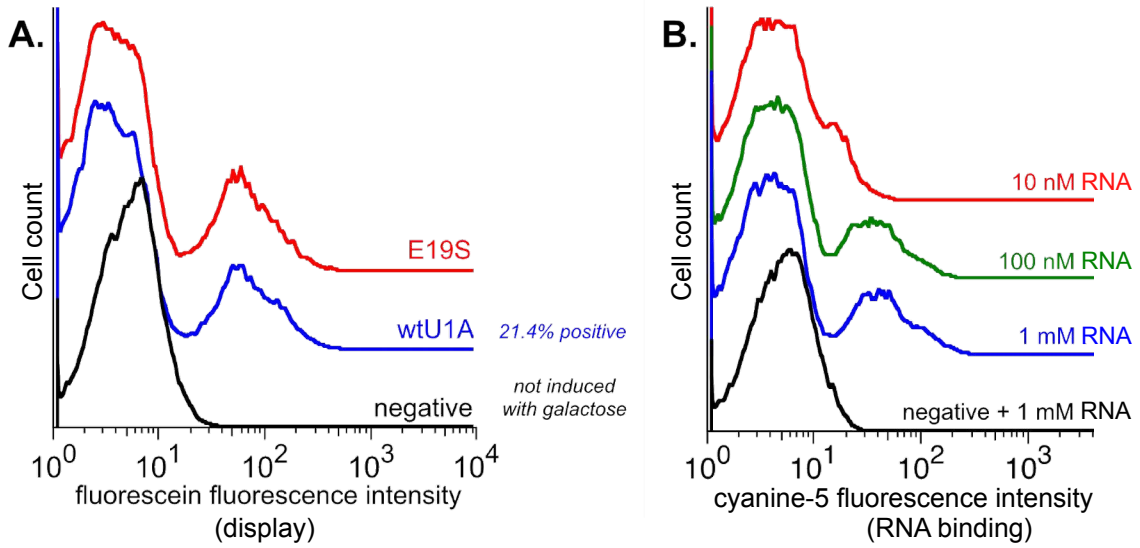


Figure 2.3: Functional Display of U1A Variants The histogram in **A** shows yeast displaying *myc* fusions of wtU1A, E19S, or displaying nothing (negative), demonstrating that U1A variants display reasonably well. The histogram in **B** shows yeast displaying wtU1A incubated with various concentrations of fluorescently labelled U1hpII, and demonstrates that the displayed protein is still functional while displayed on a yeast cell.

2.5 Library Preparation

2.5.1 Cloning the β2β3 Loop Library

As was alluded to earlier, our decision to target full sequence-space coverage of a limited region of the protein meant that we needed to perform saturation mutagenesis [111] of positions 46 and 48–51 in the β2β3 loop of U1A E19S. We chose to use NNK codons (N = A/T/C/G, K = G/T) to generate diversity. There are 4×4×2 = 32 possible NNK combinations, rather than the 4^3 = 64 possible NNN combinations. There is no loss in sequence diversity, since all 20 proteinogenic amino acids are coded for by at least one NNK codon, and only one NNK codes for "Stop" (TAG), which means fewer transformants contain truncated proteins.

The initial challenge in making our saturation mutagenesis library was the lack of convenient cloning sites near the β2β3 loop. To overcome this, we chose to use the type IIs restriction enzyme BsaI, which, when properly designed, is able to create DNA fragments which, when ligated, produce a construct without a cloning scar.

AmpR BsaI Removal

In order to use BsaI, we had to first remove the BsaI recognition site from the Ampicillin resistance region in pCTcon2 (5′-GAGACC-3′→ 5′-GAGCGC-3′), which was performed using a standard site-directed-mutagenesis protocol and the primers FWD BsaI Out (5′-CAA GGA GGT GTC GAG C GCC ACC AAC-3′) and Rev BsaI Out (5′-CTC GAC ACC TCC TTG AAG ATG ACA AAA GCT TGG CC-3′). This "BsaI out AmpR" plasmid is used for all pCTcon2 constructs moving forward. The plasmid was digested with NheI and BamHI, treated with CIP, and extracted from a 1% agarose gel after 40 minutes of electrophoresis at 140 V in TBE buffer.

1st Gen Lib Receiving Plasmid

An insert encoding, from 5′ to 3′, an NheI cut site, a BsaI restriction site, and the 3′ portion of U1A (coding for Arg61 to the C-terminus, 75 residues including the *myc* tag) was prepared via PCR using primers FWD U1A BsaI (5′-CAA GGA GGT GTC GAG C GCC ACC AAC-3′) and reverse U1A BamHI (5′-CGG GAT CCT GCG GCC GCA ACC-3′). The resulting amplicon was digested with NheI and BamHI, and ligated into the pCTcon2 plasmid described in the previous paragraph, to make *1st Gen Lib Receiving Plasmid*, also described as *BsaI U1A* plasmid henceforth.

β2β3 Loop Library Amplification and Ligation

Next, the β2β3 library was created amplifying the 5′ portion of U1A E19S (Positions 1- 53) with 5 sites in the β2β3 loop region (Ser46, Ser48, Leu49, Lys50, and Met51) substituted with NNK codons (or, in this case, the reverse complement of an NNK: MNN). An amplicon was generated via PCR using primers Fwd U1A NheI and Rev b2- b3 lib (5′-TA TAT GGT CTC GCC CCT MNN MNN MNN MNN CCG MNN TAC CAG GAT ATC CAG GAT CTG GCC-3′).

To receive this insert, *BsaI U1A* Plasmid was digested with NheI and BsaI, treated with CIP, and extracted from a 1% agarose gel to give pure vector suitable for cloning. The insert and vector were then combined using a Quick Ligation Kit (NEB) according to the manufacturer's instructions. Ligated vector was purified by phenol chloroform extraction (3x), chloroform extraction (2x), and ethanol precipitation.a

The resulting DNA was used as a template for a second PCR with homologous recombination primers for cut pCTcon2 HR FP (5′-CTC TGG TGG AGG GCG TAG CGG AGG CGG AGG GTC GGC TAG C-3′) HR RP (5′-CGA GCT ATT ACA AGT CCT CTT CAG AAA TCA GCT TTT GTT CGG ATC C-3′) (This pair of primers will be simply referred to as "HR Primers" henceforth), which are designed to create an insert with ~40 base pairs of overlap with the pCTcon2 vector, enabling yeast to perform homologous recombination on the two linear pieces of DNA, removing the need to ligate the DNA. The resulting amplicon contained the randomized β2β3 loop library in a complete U1AE19S gene ("Library Amplicon", Section C.1.5), and was purified by gel electrophoresis. The general cloning scheme is shown in Figure 2.4.

2.5.2 Library Transformation

To transform the library, pCTcon2 vector was cut with BamHI and NheI, and gel purified. Five aliquots containing ~5 µg library amplicon and ~2 µg cut pCTcon2 were prepared, and the DNA ethanol precipitated. These ethanol precipitated pellets were used to perform five separate yeast electroporations, which were each rescued in 1 mL 30 °C YPD and immediately combined. Yeast were allowed to incubate in the YPD at 30 °C for ~60 minutes, followed by a centrifugation (2500 x *g*, 5 minutes, 4 °C), removal of YPD, and resuspension in 50 mL liquid SD-CAA. Serial dilutions of this resuspension were plated onto SD-CAA agar plates, and after 3 days of growth, the number of colonies was counted to determine the number of transformants. The final determination was that there were ~6.6×10^6 transformants. While this likely does not represent a complete library, it was deemed sufficient.

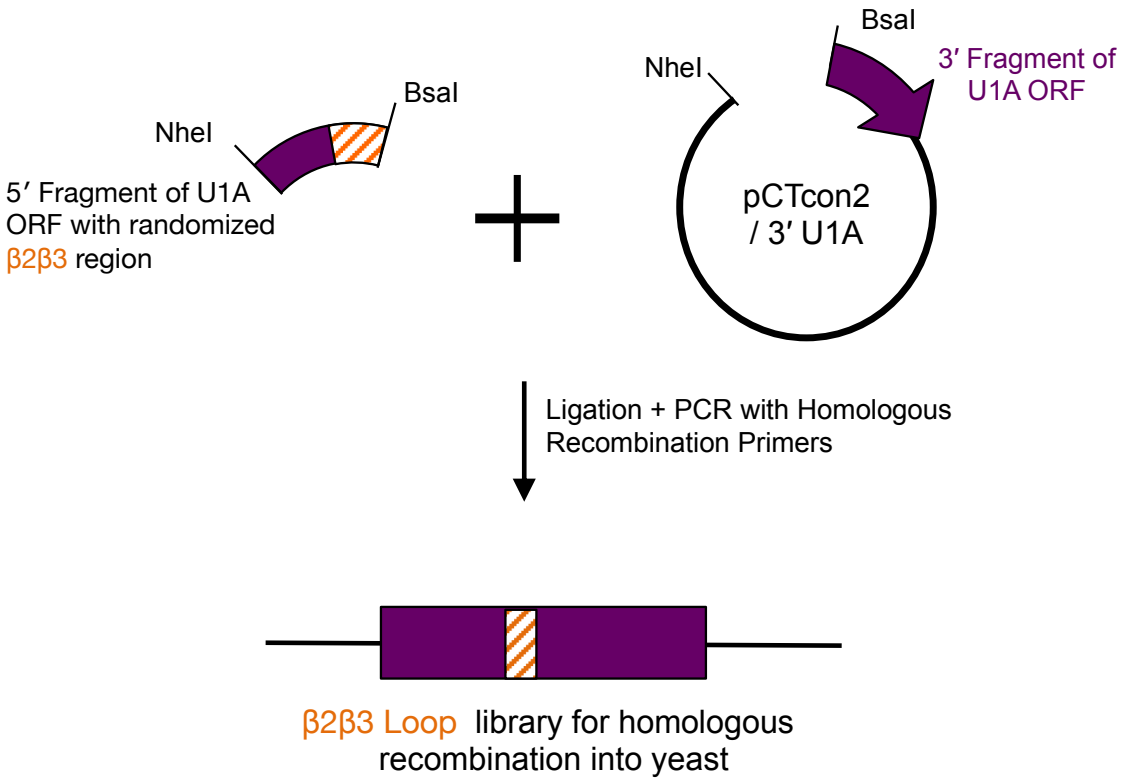


Figure 2.4: Cloning Diagram for Diversification of the β2β3 Loop

2.6 Library Screening

2.6.1 General Screening Conditions

The yeast culture containing our β2β3 loop library was prepared by standard galactose induction (subculture in SD-CAA to a density of ~0.5 × 10^7) cells/mL, grow to a density of ~2.0 cells/mL, subculture in SG-CAA to a density of ~1.0 × 10^7) cells/mL. Grow for 24–36 hours).

To prepare the library for screening the library, ~10^8 yeast cells were pelleted, washed in 500 µL ice-cold PBS-BSA. As a control, uninduced yeast cells were subjected to the same conditions as the experimental samples each round to ensure that the library was performing above background. The library screened was treated with 1:1000 FITC conjugated anti-myc antibody, as well as varying concentrations of TAR RNA labeled with a 5′ Cy-5. Prior to being used in screening the 5′ Cy-5 TAR RNA was melted and refolded (2.4.2). Varying concentrations of *E. coli* tRNAs were used as an off-target competitor. *E. coli* tRNAs were chosen as competitors due to the fact that they have ~2x the number of residues as the TAR RNA target, and these residues form a variety

of secondary structures. Presumably, any U1A variants which merely have general affinity for any RNA, or even general affinity for *structured* RNA, rather than the desired *specificity* for TAR RNA, would preferentially bind the unlabelled tRNAs.

Because FITC and Cy-5 have essentially orthogonal emission profiles it is possible to plot each flow cytometry event on a graph with FITC and Cy-5 fluorescence visualized on the X and Y axes, representing cells displaying protein and binding RNA respectively. The ideal case is that a yeast cell is both displaying a U1A variant, and that it has bound TAR RNA *because* of this U1A variant. The display can be seen with FITC fluorescence, RNA binding with Cy-5 fluorescence, and a yeast cell binding RNA *because* of its displayed protein will be "double-positive," and appear in the top-right quadrant of a yeast display plot (clearly seen in Figures 2.5 and 2.7).

2.6.2 Sorting Methods

Each round, we used a MoFlo flow cytometer to sort double positive cells into ~5 mL of SD-CAA, each time aiming to collect the best-binding ~2% of the cells displaying a U1A variant. Sorting took around 2 hours for each round. The SD-CAA the cells had been sorted into was transferred to 50 mL of SD-CAA within an hour of sort completion, and the cells allowed to grow to confluency (usually 2–3 days of growth). Upon reaching confluence, the cells were sub-cultured and induced, and the process was repeated.

After each round, we extracted plasmid DNA from an aliquot of the outgrowth using a ZymoPrep II extraction kit (Zymo Research). This DNA was transformed into 5α *E. coli*, and ~10 sequences were analyzed. No obvious pattern arose in these first three rounds, and this sequencing was simply to ensure that we had not *already* converged upon a sequence. The sequencing data can be seen in Table 2.1.

2.6.3 Sorting Conditions

Conditions became more stringent as the selection was continued based on a combination of reducing Cy-5 TAR concentration or incubation time, while raising either incubation temperature or competitor tRNA concentration. Rounds one and two were performed at 25 °C with 5 μM

Table 2.1: Sequences of library members in first three rounds of sorting

Position	46	48	49	50	51
wt	*Ser*	*Ser*	*Leu*	*Lys*	*Met*
			Round 1		
1.1	R	R	R	S	Q
1.2	S	W	T	L	A
1.3	V	H	G	F	A
1.4	P	M	R	R	L
1.5	S	F	T	P	P
1.6	S	L	L	T	D
1.7	A	H	K	V	S
1.8	K	C	V	F	S
			Round 2		
2.1	C	P	C	T	Y
2.2	T	N	Y	T	F
2.3	G	P	S	P	H
2.4	V	L	D	Y	T
2.5	Q	L	*	A	M
2.6	S	P	*	V	S
			Round 3		
3.1	C	P	C	R	R
3.2	V	L	A	S	C
3.3	P	P	R	R	P
3.4	G	R	R	C	T
3.5	T	Q	G	L	K
3.6	C	C	A	K	N
3.7	L	L	V	A	S
3.8	S	S	H	Q	A

unlabeled *E. coli* tRNAs, while round three was performed at physiological temperature (37 °C) with 50 μM unlabeled tRNAs from E. coli. The concentration of TAR-Cy5 (round 1: 10 μM; round 2: 1 μM; and round 3: 0.5 μM) and incubation time (round 1: 60 min; round 2: 30 min; and round 3: 30 min) were also decreased as rounds increased to enrich the highest affinity interactions.

The variation in conditions over the six rounds screening is shown in Table 2.2. The fluorescence profiles of the populations in each round, along with approximations of which cells were collected, are shown in Figure 2.5. Rounds 4–6 occur after the diversification discussed in Section 2.6.4, and represent the C-Helix library.

Table 2.2: Yeast Display Round Conditions

Conditions	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6
[TAR-Cy5, nM]	10,000	1,000	500	500	100	10
[tRNA, nM]	4,000	4,000	50,000	50,000	50,000	50,000
incubation time (min)	60	30	30	30	30	30
incubation temp (°C)	25	25	37	37	37	37
yeast screened	$\sim 4 \times 10^7$	$\sim 6 \times 10^7$	$\sim 3 \times 10^7$	$\sim 3 \times 10^7$	$\sim 3 \times 10^7$	$\sim 1 \times 10^7$
yeast sorted	$\sim 1 \times 10^6$	$\sim 6 \times 10^5$	$\sim 2 \times 10^5$	$\sim 5 \times 10^5$	$\sim 2 \times 10^6$	$\sim 3 \times 10^5$

2.6.4 Diversification

The only change in this routine came after round 3, when the already sorted β2β3 loop library was diversified by adding randomized sequences to an additional region important for binding. The diversification was performed at one of two locations—The β1α1 loop (positions 15, 16, and 19), or the C-helix (positions 91, 92, and 94). Since each library has 3 randomized codons, they represent an 8000-fold increase in diversity (20^3) from the sorted library of unknown size. A diagram of this diversity can be seen in Figure 2.6.

As before, we used BsaI to ensure that there were no cloning scars. A modified pCTcon2 plasmid, named "2nd Gen Lib Recieving Plasmid"(Sequence: Section C.1.6) , was constructed to receive the second-generation library using the following methods. A pCTcon2 plasmid containing *wtU1A* downstream of Aga2 and a linker was digested with restriction enzymes MluI and NotI

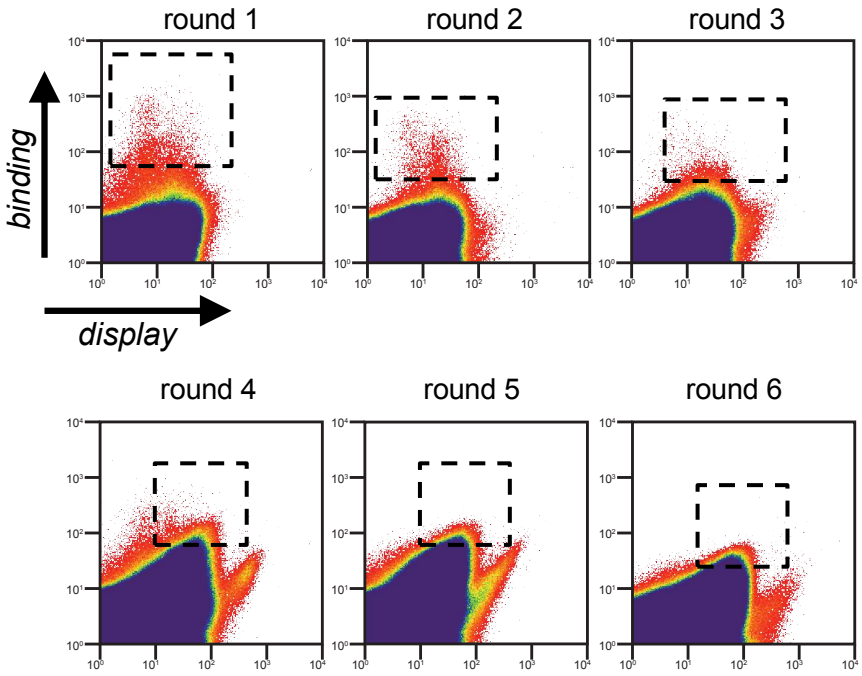


Figure 2.5: Yeast Sorted in Rounds 1–6 An approximation of yeast cells display and binding, along with an approximation of which cells were sorted for the next round. On these two-color graphs cooler colors represent a higher density of events at that position. We aimed to take ~2% of the yeast displaying an RRM in each round. Adapted from [109].

(both from NEB), followed by treatment with Calf Intestine Phosphatase (NEB), and extracted from a 1% agarose gel to give pure vector suitable for cloning. Then, two oligos, Fwd c-helix recieving (5′-CGC GTC CTA ACC ACA CTA TTT ATA TGA GAC CAC TCT AGA GGT TCC CCG GTT GC-3′) and Rev c-helix receiving (5′-GGC CGC AAC CGG GGA ACC TCT AGA GTG GTC TCA TAT AAA TAG TGT GGT TAG GA-3′), were treated with T4 Polynucleotide Kinase (NEB) according to the manufacturer's instructions. The phosphorylated oligos were then heated up to 94 °C for 5 minutes and allowed to room temperature slowly, over 5 minutes. The annealed oligos were then cloned into the vector using standard methods to create *Dual BsaI U1A pCTcon2*.

The C-helix library was built by using the DNA extracted from the 3rd round sort and introducing diversity at the C-helix position on the U1A scaffold, while maintaining the selected diverse β2β3 loops. The U1A mutant genes were amplified by PCR using the primers Fwd U1A NheI and Rev c-helix lib (5′-T ATA TGG TCT CTT GGC MNN GAT MNN MNN GTC GGT GCG CGC AT ACT GGA TAC G- 3′). The resulting amplicon was digested with NheI and BsaI.

To receive this insert, *2nd Gen Receiving Plasmid* was also digested with NheI and BsaI, and treated with Calf Intestine Phosphatase (NEB), and extracted from a 1% agarose gel to give pure vector suitable for cloning. The insert and vector were then combined using a Quick Ligation Kit (NEB) according to the manufacturer's instructions. Ligated vector was purified by phenol chloroform extraction (3x), chloroform extraction (2x), and ethanol precipitation. The resulting DNA was used as a template for a second PCR with HR Primers. The resulting amplicon contained the β2β3 loop sequences isolated from three rounds of screening with a randomized C-helix region (Sequence: Section C.1.7), and was electroporated into EBY100.

The β1α1 library was produced using similar methods, but the receiving plasmid was constructed to contain a BsaI site upstream of the β1α1 loop sequence. The library was amplified Rev U1A BamHI primer and a forward primer which randomized the β1α1 region, Fwd b1a1 (5′-ATA TAG GTC TCT TAT ATC NNK NNK CTC AAT NNK AAG ATC AAG AAG GAT GAG CTC AAA AAG-3′). The recieving plasmid and amplicon were digested with BsaI and *BamHI*, and the insert ligated in. All other cloning methods were identical.

Again, a visual representation of the diversification process can be seen in Figure 2.6.

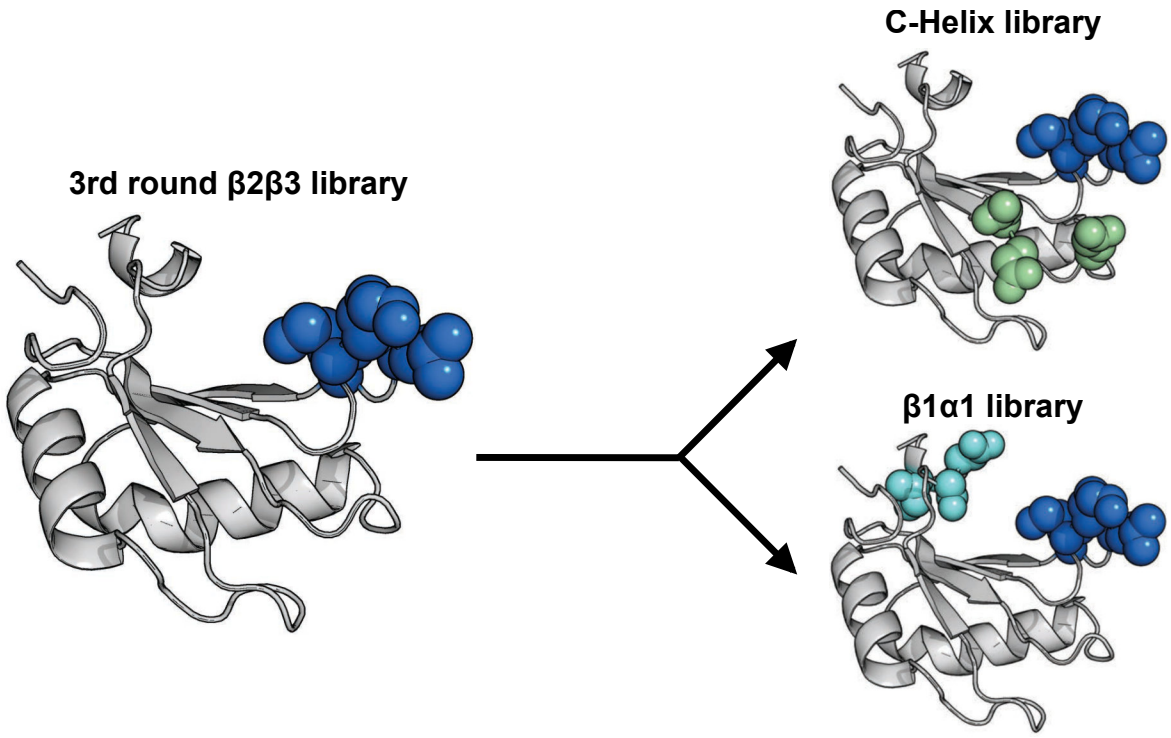


Figure 2.6: Diversification Strategy Having narrowed down our randomized β2β3 loop, we aimed to diversify our library at one of areas, but not both simultaneously. Adapted from [114]

The C-Helix library yielded ~2 × 10^7, and the β1α1 library ~3 × 10^7 transformants. Though there was no way to *know* how many members a "complete" library contained at this stage, this was deemed sufficient. Both libraries represent an increase in diversity of ~8000-fold (20^3), so this represents a complete library if we had ~400 β2β3 sequences remaining.

2.6.5 Screening of Rounds 4–6

Rounds 4–6 were performed as rounds 1–3 were, with the conditions given in Table 2.2. Gratifyingly the C-Helix library represented a major improvement in double positive population from the previous round, despite the now rather stringent sorting conditions with 50 μM *E. coli* tRNA acting as competitor. As can be seen in Figure 2.7, the C-Helix library significantly outperformed the β1α1 library in both rounds 4 and 5, which led to the abandonment of the β1α1 library.

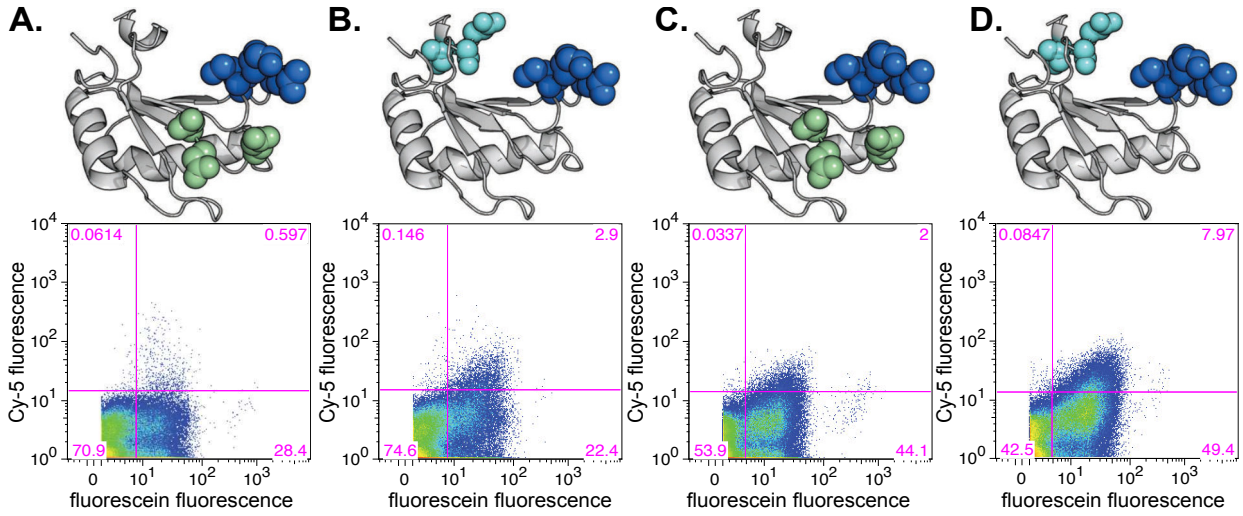


Figure 2.7: Yeast Display for Rounds 4 and 5 We analyzed both the randomized C-helix and β1α1 libraries in rounds 4 and 5. **A** shows the β1α1 library and **B** the C-helix library under round 4 conditions while **C** shows the β1α1 library and **D** shows the C-helix library under round 5 conditions (conditions shown in Table 2.2. As can be seen, the C-Helix library was superior under both attempts. Adapted from [114])

2.6.6 Sequence Analysis of Round 4

We analyzed sequencing data for both the C-helix and the β1α1 libraries after round 4 (though not round 5), these sequences are shown in Table 2.3. Though no obvious patterns emerged, there are some interesting observations to be made in hindsight. The first is the abnormally high incidence of proline at positions 46 and 51, which would become a signature of the final library. The other is that even the β1α1 fourth generation library seems to be circling the ultimate consensus sequence, with sequence 4.a.8 actually *having* the β2β3 loop consensus sequence.

2.6.7 C-Helix Rationale

I hypothesize that the ultimate success of the C-Helix library over the β1α1 library, despite the apparent similarity in β2β3 loop sequences, is due to the fact that the C-Helix is the most dynamic piece of the RRM. In acting as a mobile clamp, it is likely that if it is unable to form favorable interactions, it simply doesn't participate at all. Given what we would come to learn about this binding interaction (described in Chapter 4), it is frankly possible that we were just giving it the ability to *not* participate in the binding interaction, and therefore removing possible steric interference. It is also possible that the dynamism of this element means there are more

Table 2.3: Yeast sorted in fourth round, both from the C-Helix and β1α1 libraries

	β2β3 Loop Position					C-Helix Position		
	46	48	49	50	51	91	92	94
wt	*Ser*	*Ser*	*Leu*	*Lys*	*Met*	*Ser*	*Asp*	*Ile*
4.1	P	K	R	T	P	N	A	Q
4.2	*	P	E	Q	G	Y	M	F
4.3	P	R	R	Q	P	E	S	P
4.4	P	T	R	L	P	S	G	E
4.5	S	C	I	I	P	L	Y	*
4.6	P	A	R	R	F	P	E	P
4.7	R	H	R	R	P	E	P	D
4.8	A	R	A	S	R	T	A	S

	β2β3 Loop Position					β1α1 Position		
	46	48	49	50	51	15	16	19
wt	*Ser*	*Ser*	*Leu*	*Lys*	*Met*	*Asn*	*Asn*	*Ser*
4.a.1	P	T	R	R	P	S	G	A
4.a.2	E	M	I	L	C	R	Y	G
4.a.3	K	Y	R	T	P	S	G	A
4.a.4	P	S	R	R	P	S	Y	G
4.a.5	T	R	M	R	L	G	G	S
4.a.6	P	S	R	R	P	G	H	T
4.a.7	P	D	I	N	R	P	R	I
4.a.8	P	T	R	T	P	G	Q	Y

"favorable" possibilities, rather than requiring a specific arrangement (synergistic with the β2β3 loop) to get a "*maximally* favorable" possibility. In U1hpII, the backbones of S91, D92, and I94 form a complex H-bond network with the A and C residues at the 6th and 7th positions of the U1hpII loop [110]. Notably, the TAR loop is 6 bases long, indicating that the C-Helix may be able to engage with this loop. I also posit that in being mobile, the C-helix may be more modular and independent of the interactions between β2β3 loop and TAR RNA, while the β1α1 loop is likely to require synergistic maturation concomitantly with the β2β3 loop.

2.7 Properties of Sixth Generation TAR Library

2.7.1 β2β3 Loop Homology

After six rounds of sorting, U1A E19S derived sequences for the sixth generation were analyzed. Since these were the 6th generation of TAR Binding Protein, a designation of "TBP 6.X" was given to each protein. A total of 71 sequences were analyzed, and all sequences are included in Table 2.4.

The sequence logo for the β2β3 loop can be seen in Figure 2.8A.

This logo shows a strong consensus sequence for the β2β3 loop, suggesting that we had evolved a privileged sequence for TAR recognition. The overall charge character of the consensus sequence β2β3 loop in positions 46–51—**PRTRTP** (with R47 unmutated)—is cationic, as RNA-binding proteins are wont to be, but does not have the overwhelming positive charge which would indicate a non-specific binder. The only position which changed from non-cationic to cationic is the leucine→arginine mutation at position 48.

Positions 46 and 51, on the other hand show a very strong (100 and 93%, respectively) preference for proline—a neutrally charged, somewhat hydrophobic residue notable for major *structural* effects—in contrast to the serine and (cationic) lysine present at these positions in the wild-type U1A protein. The structural consequences of these flanking prolines became clear with the crystal structure (Figure 4.4), but in the moment the most important conclusion was that we *weren't* merely selecting for positive charge. Position 48 is largely populated by threonine and

Table 2.4: Sequences from TAR 6G Library

β2β3 Loop Position	46	48	49	50	51	**C-Helix Position**	91	92	94
wtU1A	*Ser*	*Ser*	*Leu*	*Lys*	*Met*		*Ser*	*Asp*	*Ile*
TBP 6.1	P	T	R	T	P		P	P	P
TBP 6.2	P	T	R	T	P		A	R	K
TBP 6.3	P	T	R	R	P		R	T	R
TBP 6.4	P	T	R	T	P		K	H	I
TBP 6.5	P	R	R	T	W		R	H	Q
TBP 6.6	**P**	**T**	**R**	**T**	**P**		**G**	**R**	**A**
TBP 6.7	**P**	**Q**	**R**	**T**	**P**		**K**	**R**	**P**
TBP 6.8	P	T	R	T	P		D	R	T
TBP 6.9	P	R	R	T	P		R	T	K
TBP 6.10	P	T	R	R	P		G	R	R
TBP 6.11	P	T	R	T	P		S	Q	P
TBP 6.12	P	T	R	T	P		S	R	G
TBP 6.13	P	T	R	R	P		R	R	P
TBP 6.14	P	T	R	T	P		V	P	V
TBP 6.15	P	R	R	T	P		R	P	P
TBP 6.16	P	T	R	N	P		T	K	A
TBP 6.17	P	R	R	T	Y		A	P	K
TBP 6.18	P	T	R	T	P		S	K	P
TBP 6.19	P	T	R	T	P		D	K	R
TBP 6.20	P	R	R	T	P		R	P	K
TBP 6.21	P	T	R	T	P		T	K	P
TBP 6.22	P	T	R	T	P		P	R	P
TBP 6.23	P	Y	R	T	P		R	R	A
TBP 6.24	P	T	R	T	P		G	K	R
TBP 6.25	P	R	R	T	P		T	N	K
TBP 6.26	P	R	R	T	P		S	A	V
TBP 6.27	P	T	R	R	P		S	A	R
TBP 6.28	P	P	R	R	P		R	P	R
TBP 6.29	P	R	R	T	Y		S	R	P
TBP 6.30	P	T	R	T	P		S	R	P
TBP 6.31	P	T	R	T	P		R	P	S
TBP 6.32	P	M	R	R	P		S	H	Q
TBP 6.33	P	T	R	T	P		K	C	P
TBP 6.34	P	R	R	T	P		R	P	Q
TBP 6.35	P	T	R	T	P		S	R	V

β2β3 Loop Position	46	48	49	50	51	**C-Helix Position**	91	92	94
wtU1A	*Ser*	*Ser*	*Leu*	*Lys*	*Met*		*Ser*	*Asp*	*Ile*
TBP 6.36	P	R	R	T	P		L	L	P
TBP 6.37	P	R	R	T	P		G	K	R
TBP 6.38	P	T	R	T	P		L	R	R
TBP 6.39	P	T	R	T	P		R	Q	R
TBP 6.40	P	T	R	V	P		A	R	W
TBP 6.41	P	T	R	T	P		G	P	P
TBP 6.42	P	T	R	T	P		E	A	P
TBP 6.43	P	P	R	T	Y		L	I	Q
TBP 6.44	P	K	R	T	P		L	V	P
TBP 6.45	P	T	R	T	P		K	T	S
TBP 6.46	P	M	R	T	P		G	R	A
TBP 6.47	P	T	R	T	P		W	A	P
TBP 6.48	P	T	R	T	P		G	R	S
TBP 6.49	P	R	R	T	Y		T	R	K
TBP 6.50	P	T	R	T	P		K	P	P
TBP 6.51	P	R	R	Q	P		H	R	R
TBP 6.52	P	Y	R	T	P		P	P	R
TBP 6.53	P	R	R	T	P		T	N	K
TBP 6.54	P	Q	R	T	P		R	R	S
TBP 6.55	P	R	R	T	P		G	W	K
TBP 6.56	P	H	R	T	P		G	R	Q
TBP 6.57	P	T	R	T	P		A	R	A
TBP 6.58	P	T	R	T	P		S	K	R
TBP 6.59	P	R	R	T	P		T	P	T
TBP 6.60	P	T	R	T	P		G	K	R
TBP 6.61	P	P	R	T	P		A	H	T
TBP 6.62	P	R	R	T	P		R	T	P
TBP 6.63	P	T	R	T	P		S	Q	L
TBP 6.64	P	T	R	T	P		R	A	S
TBP 6.65	P	T	R	T	P		S	G	P
TBP 6.66	P	T	R	T	P		S	P	G
TBP 6.67	P	T	R	V	P		Q	R	N
TBP 6.68	P	T	R	T	P		T	P	G
TBP 6.69	P	R	R	T	P		G	K	P
TBP 6.70	P	H	R	T	P		G	R	A
TBP 6.71	P	R	R	T	P		L	R	P

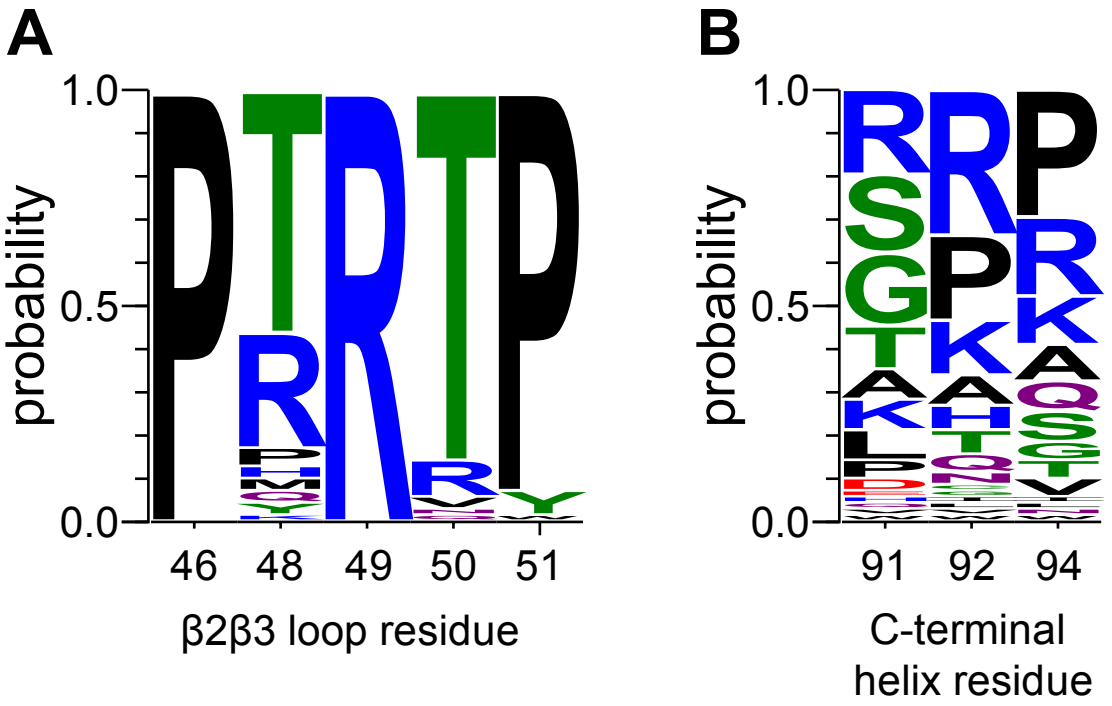


Figure 2.8: TAR 6G Sequence Logo The sequence logo shows that **A** the β2β3 loop showed excellent homology after sorting and **B** the C-Helix region showed patterns, but no homology

arginine, which means that both the wt and library consensus sequence have an alcohol functionality in position 48 (serine in the wild-type and threonine in the consensus). But there are two instances of asparagine among the 71 sequences analyzed, one of them in TBP 6.7, the most avid TAR binding protein we characterized.

2.7.2 C-Helix Homology

The sequence logo for the C-helix, shown in Figure 2.8B, shows that this region is considerably more heterogeneous than the evolved β2β3 loop, but certain themes emerge. In native U1A, residues 91 and 92 are serine and aspartic acid, respectively. Interestingly, a signficant portion of the TAR-binding RRMs retain serine at position 91, although arginine, glycine, and threonine are also prominent. Positions 92 and 94 undergo more dramatic changes. In particular, cationic residues arginine and lysine are found in a significant number of TAR-binding RRMs, as is proline. Proline, arginine, and lysine are also prominent at position 94, which in the native protein is known to modulate the conformation of the C-terminal helix. The prominence of proline at

positions 92 and 94 within the C-terminal helix is interesting and suggests that conformational rigidity might be favorable at this position, while the fairly prominent glycine at position 91 may indicate that flexibility may be more favorable in that position. In any case, we would eventually learn that the C-Helix region was inessential for TAR binding [45], which seemed to contradict the apparent variation in affinity *based* on different C-Helix sequences shown in Figure 3.3. The most plausible explanation is that we simply selected for a C-Helix which stays out of the way of the β2β3 loop, and doesn't interfere with bindingrawford.

2.8 Initial Characterization Attempts

Our goal of determining quantitative binding values (i.e. dissociation constants, which will be expressed generally as K_D values in this work) proved challenging. Initially we tried to determine a K_D via Fluorescence Polarization (as had been done for U1A E19S) and Yeast Display. Both failed to give believable and reliable data, and initial assays were a mix of disheartening and promising.

2.8.1 Fluorescence Polarization

Fluorescence Polarization had been used to good effect to quantify the binding of the original engineered synthetic RRMs in the McNaughton Lab, most notably the U1A E19S which had been the starting point for our scaffold. As such, it seemed a logical choice to analyze these new putative TAR binders.

Briefly, the protein being analyzed was concentrated to ~1 mM and used to load on a plate, and was diluted until its concentration ranging from the maximum ~1 mM to ~1 nM. These dilutions were added to a 384 well plate, and a fixed concentration of RNA (melted and refolded 2.4.2) was added to each well with a final RNA concentration of 20 nM. Full method given below.

Fluorescence Polarization Method

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 protein

concentrations. The appropriate fluorescein tagged nucleic acid (Integrated DNA Technologies, RNase free HPLC purified and shipped as a lyophilized pellet) 110 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.

Results

We analyzed a member of the pre-diversification from after the 3rd sort, and a library member from after the 6th sort (TPB 6.2, which would prove to be a middle-of-the-range binder in later assays). The results, which can be seen in Figure 2.9 were discouraging. TBP 3.1 had an apparent K_D of ~30 μM, 6-fold worse than our starting point, while TBP 6.2 an apparent K_D of ~5 μM, no better than our starting point. With that said, since neither curve seemed to saturate, the conclusion was that this assay was fundamentally unfit to analyze these particular protein–RNA interactions.

2.8.2 Characterization via Qualitative Yeast Display

Though it did not ultimately prove to be the best quantitative way to analyze our TBP 6.X proteins, we had more encouraging results analyzing our TBP 6.X proteins via yeast display. As was discussed in Section 2.3.1, the ability of yeast display to be used as a standalone analytical technique is a major advantage to using it as a screening platform. We were simply able to electroporate the isolated plasmid we had used to analyze our sequence back into EBY 100 yeast, and induce display as we had during our controls (Section 2.4.2).

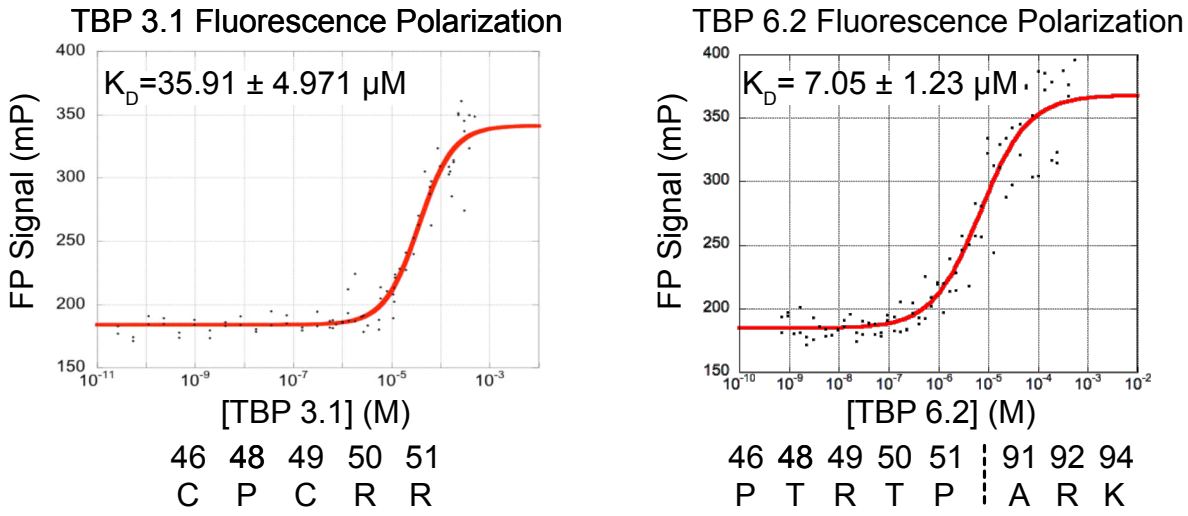


Figure 2.9: TAR Binding Protein 3.1 and 6.2 Fluorescence Polarization We attempted to use Fluorescence Polarization to characterize our library, but did not see any saturation. These data were worrying, because they indicate no improvement in TAR binding

TAR 6G.XX proteins were analyzed in two batches under two different sets of conditions in order to give two subtly different pieces of comparison data. The first batch was measured by their absolute ability to bind TAR RNA while displaying on a yeast cell. The samples were incubated under the same conditions as the 6th generation sort (10^8 cells, 10 nM labeled TAR, 1:1000 FITC conjugated anti-myc antibody, 37 °C, 30 minutes). The results are shown in Figure 2.10A. A negative sample under these conditions has >99% of events within the "no display, no binding" quadrant in the bottom left.

For all but one sample (TBP 6.5), there was clear evidence of RNA binding, even at the relatively low concentrations of TAR we used. However, there was fairly high variations in the amount of display, making this a difficult technique to apply quantitatively.

Since the ultimate goal was to make a *better* binder of TAR RNA than U1A E19S, I next took a series of samples with the explicit goal of comparing their TAR binding to that of U1A E19S. These were prepared with 6th round conditions, but without any FITC conjugated anti-myc antibody, since I was only looking at RNA binding. We found that U1A E19S displayed on TAR generates only a very small positive signal at these conditions, but increases somewhat when the TAR RNA concentration is increased to 10 μM (around the putative K_D of U1A E19S).

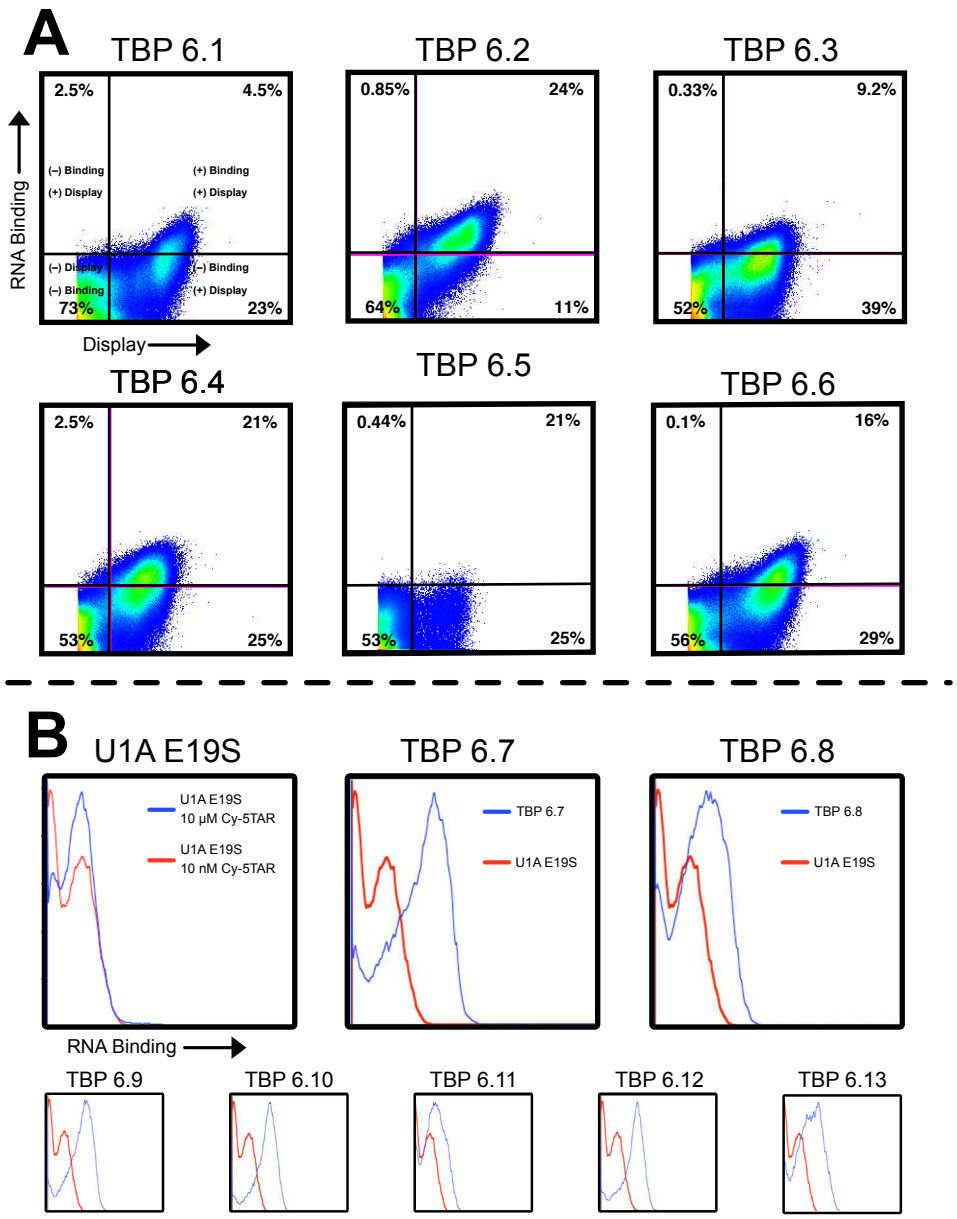


Figure 2.10: Qualitative Yeast Display for Characterizing Generation 6 TAR Binding Proteins Shown in **A**, an analysis of our library under round 6 conditions showed that there was promising binding activity in an absolute sense, and **B** A comparison of the RNA binding activity of our 6th generation library to the U1A E19S starting point indicated that our new TAR binders were more avid binders of RNA than was our starting material

We also found that our TBP 6.X variants had both more Cy-5 positive events, and that these Cy-5 positive events were brighter (which indicates *more* RNA binding on a given yeast cell). Examples are shown in Figure 2.10B. Though the RNA binding analysis was qualititative rather than quantitative, and lacking display information, was imperfectly controlled, it still indicated that our members of our 6th generation library bound TAR better than the U1A E19S starting material did.

2.8.3 Yeast Display K_D

Between the two rounds of qualitative yeast display, we attempted (with the help of our labmate Bryce Rogers) to perform a *quantitative* yeast display assay. It is possible to get a K_D from a yeast display experiment [113] by measuring multiple samples with varying concentrations of target, and plotting the *mean* fluorescence (rather than the percent positive) against concentration of the RNA used in the analysis. In this case, the yeast were incubated with varying amounts of TAR RNA ranging from 1 nM to 1 µM. We chose two variations—TBP 6.2 and TBP 6.4. These data are shown in Figure 2.11.

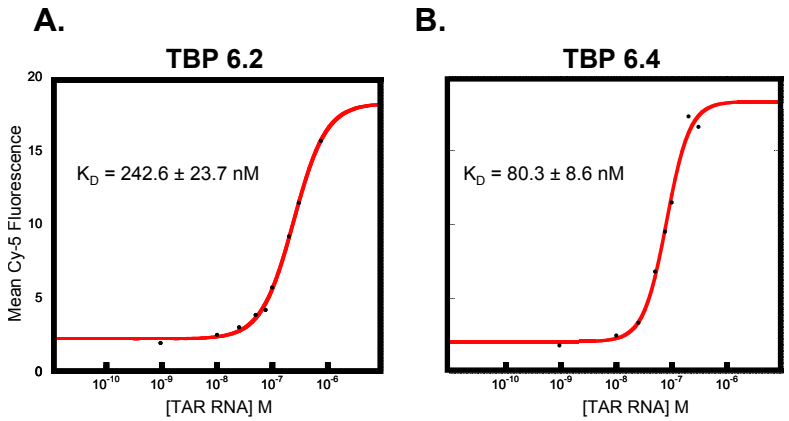


Figure 2.11: Quantitative Yeast Display Assays of TBP 6.2 and 6.4 The K_D values estimated by a quantitative yeast display were promising. Both seemed to be ~100 nM, which would represent two orders of magnitude improvement over our ~10 µM starting point. Red curve represents the best fit of the Hill Equation to these data using Kaleidagraph.

Though this was not perfectly convincing, since the mean fluorescence signal never reaches a saturation point, fitting the Hill Equation yielded values of ~250 nM and ~80 nM respectively for TBP 6.2 and 6.4 respectively, with reasonable error. Though not *reliable*, since the signals did not saturate, this was encouraging, as it would represent a full order of magnitude improvement over the putative K_D determined via Fluorescence Polarization (setting aside questions of comparing K_D values derived from different assays).

Though both K_Ds were, at this point putative, we had already performed the qualitative yeast display from Figure 2.10A that showed excellent absolute binding, and I would soon gather the data shown in Figure 2.10B, which would show a significant degree of improvement over our U1A E19S starting point. As such, in the months in which we lacked a reliable quantitative binding value, these ~100 nM K_D values (Figure 2.11) were thought to be more reflective of reality than the ~10 μM K_D values from the Fluorescence Polarization assay (Figure 2.9).

2.9 Conclusions

Though it would not become clear until later, the screen described in this chapter was a great success. For me personally, the work described in this chapter represents the time I spent learning to be a competent molecular biologist, both technically and in being able to design and perform experiments. The foundation for much of the rest of my time doing graduate research was laid here.

For the project at large, the fact is that the reason that our initial characterizations failed was that the interaction being analyzed was *tighter* than permissible by the assays we were using to analyze them, and this would lead me to perform ELISA assays (shown in Section 3.3) with a precision that I initially didn't think was possible, and lead us to characterize the tightest binders of TAR yet known in the work described in Chapter 3.

Chapter 3

Characterization of TAR Binding Proteins

3.1 Chapter 3 Introduction

3.1.1 Chapter 3 Summary

Having putatively demonstrated that our 6th generation library contained excellent binders of TAR RNA, we next needed to establish a convincing and quantitative basis for this claim. An ELISA protocol was determined to be a good basis for analytical comparison of library members. When applied stringently to our best binders (TBP 6.6 and TBP 6.7), it was quantitative enough to warrant proposing a K_D in the single-digit nM range. A collaboration with the Laird-Offringa lab, the most-established quantifiers of protein–RNA interactions, this K_D was shown to be in the range of 500 pM–1 nM. Further characterization established that the best proteins were quite specific for TAR RNA, and minor changes in the TAR element could abolish binding. A collaboration with the Le Grice lab enabled us observe, via SHAPE, the long-range effects this TBP 6.7 interaction had on the HIV-1 5′ untranslated region. Furthermore, our best binder, TBP 6.7, was shown by ITC to be able to inhibit the tat/TAR interaction, and via an *in vitro* transcription assay, was shown to prevent the resulting transcription cascade.

3.1.2 Chapter 3 Attribution

This chapter is adapted from [109].[2]

I was responsible for the design and performance of the ELISA screen, the quantitative K_D curves, and the analysis of the affinity of various RRMs for variants of TAR RNA (Figures 3.3, 3.4, 3.5, 3.6, 3.7, and 3.9), the ITC analysis (Figure 3.15, with assistance from Alex Chapman, also a graduate student in the McNaughton Lab), the *in vitro* transcription assays (Figure 3.17), with

[2]Crawford, DW, Blakeley, BD, Chen, PH et al. An Evolved RNA Recognition Motif That Suppresses HIV-1 Tat/TAR-Dependent Transcription. *ACS Chemical Biology*, 11(8):2206–2215, **2016**

assistance from John Anderson of the Wilusz lab in the Dept. of Microbiology at Colorado State University).

Po-Han Chen, of the Laird-Offringa Lab at the University of Southern California Dept. of Biochemistry and Molecular Biology, was responsible for the Surface Plasmon Resonance characterization.

SHAPE analysis was performed by Chringma Sherpa, working under Stuart Le Grice at the National Cancer Institute. Though I was not involved in the performance of these assays, I was frequently involved in discussing the conclusions, and they add a practical dimension to the work I performed.

3.1.3 Chapter 3 Background

The work done in Chapter 2 using yeast display to find U1A-derived binders of TAR RNA was apparently successful. The final rounds of screening indicated good amounts of TAR-binding yeast even with low concentrations of TAR, and qualitative assays indicated that certain members of our sorted library bound TAR far better than the E19S starting point. What we lacked was quantitative binding data, and functional data indicating that our TAR binding proteins could disrupt the Tat–TAR interaction.

3.2 ELISA Preparation

3.2.1 Assay preliminaries

I decided that an ELISA assay represented the best possibility of successfully developing a method to quantifiably compare TAR Binding proteins against each other, since ELISA assays are well-characterized, and use a minimum amount of material. The ELISA format favored by our lab involved using a streptavidin coated plate, and an anti-FLAG antibody with a conjugated Horseradish peroxidase (HRP) enzyme.

Since, by definition, fewer materials are used for the immobilized phase of an ELISA experiment, I decided that the easiest element to immobilize would be the TAR RNA. I purchased

this RNA directly from Integrated DNA Technologies (IDT), and it is quite expensive. This also worked well, since adding a 5′-biotin modification did not add significant cost, and would allow the RNA to be immobilized on the streptavidin coated plates commonly used in ELISA.

In order to measure the bindinginteraction, the first task, therefore, was obtaining the TBP 6.X library members described in Section 2.8.2 as purified proteins. The genes encoding these proteins needed to be transferred to a purification vector with a His_6 tag for purification, and a FLAG (DYKDDDDK) tag for analysis.

3.2.2 Cloning

I used the pETduet plasmid (digested to only contain a single T7 promoter/terminator) favored by our lab. I decided to locate fuse both the His_6 and th FLAG tag to the C-terminal of the protein. The rationale for a C-terminal His_6 tag was largely inertia: the U1A variants characterized by Brett Blakeley [105, 107] had His_6 tags located on their C-termini, and had purified well. This also seemed a logical location for the FLAG tag for steric reasons. Since the stem-loop would be the moiety of TAR furthest from the streptavidin coated plates, and the assumed binding conformation involved the RRM motif contacting the TAR loop with the C-terminal clamped above the RNA, a C-terminal FLAG tag would likely be more sterically accessible.

The genes were PCR amplified from the pCTcon2 template using U1A NcoI FP (5′-ATA TAC CAT GGC CCA GGT GCA GC-3′) and U1A His FLAG PacI RP (5′-GTT AAT TAA CTA TTA CTT GTC GTC ATC GTC TTT GTA GTC GTG ATG ATG GTG ATG ATG TGC GGC CGC AAC C-3′). This PCR product was digested with NcoI and PacI. To prepare the vector, pETduet was digested with NcoI and PacI, and treated with CIP. A 5-fold molar excess of insert was ligated into cut vector, transformed into NEB 5α, and plated. A colony was picked into 5 mL liquid culture the next day, DNA extracted, and sent for sequencing verification (GeneWiz). Protein and DNA sequences for a generic TBP 6.X protein, and all TBP 6.6 and 6.7 constructs used and DNA sequence can be found in Sections C.2.2–C.2.4.

3.2.3 Protein Purification

Upon sequence verification, plasmid was transformed into NEB chemically competent BL21 cells (C2527), and plated on LB-Carb. After overnight growth, a single colony was picked into 2 x 5 mL of culture. After overnight growth, DNA was extracted from one culture to be used as a final sequence verification.

The other culture was used as a starter culture (100X) for protein purification. Cells were grown in LB-Carb (100 μg/mL carbenicillin) in either 100 mL cultures grown in 250 mL baffled flasks, or 500 mL cultures grown in 1 L Erlenmeyer Flasks. Cells were grown at 37 °C to OD_{600} = ~0.6, when they were induced with 1 mM IPTG. Upon induction, they were incubated at 25 °C for 4–12 hours.

Cells were subsequently collected via centrifugation (5000 × *g*, 10 min, 4 °C), resuspended in 30 mL HEPES buffer (10 mM HEPES, pH = 7.4, 50 mM KCl, 30 mM NaCl, 1 mM $MgCl_2$, 1 mM EDTA) containing cOMPLETE protease inhibitor (1 tablet / 30 mL), and frozen (generally overnight) at -20 °C. Frozen cell suspensions were thawed, and sonicated for 2 minutes. The lysate was cleared via centrifugation (15000 x *g*, 25 min, 4 °C), and the supernatant mixed with 400 μL Ni-NTA agarose resin for 10 minutes (tumbling, 4 °C), and the resin collected by centrifugation (5000 x *g*, 10 minutes, 4 °C).

The resin was washed with 30 mL of HEPES buffer containing 20 mM imidazole followed by 10 mL of HEPES buffer containing 50 mM imidazole. Proteins were eluted using 4 mL of HEPES buffer containing 400 mM imidazole.

Eluted proteins were dialyzed in SnakeSkin Dialysis Tubing with a 10 kDa molecular weight cutoff (ThermoFisher 88243) against 2 L of HEPES buffer, and subsequently dialyzed against 2 L of phosphate buffer (20 mM phosphate, pH = 7.4, 150 mM NaCl).

Purified proteins can be seen cleanly on a PAGE gel in Figure 3.1.

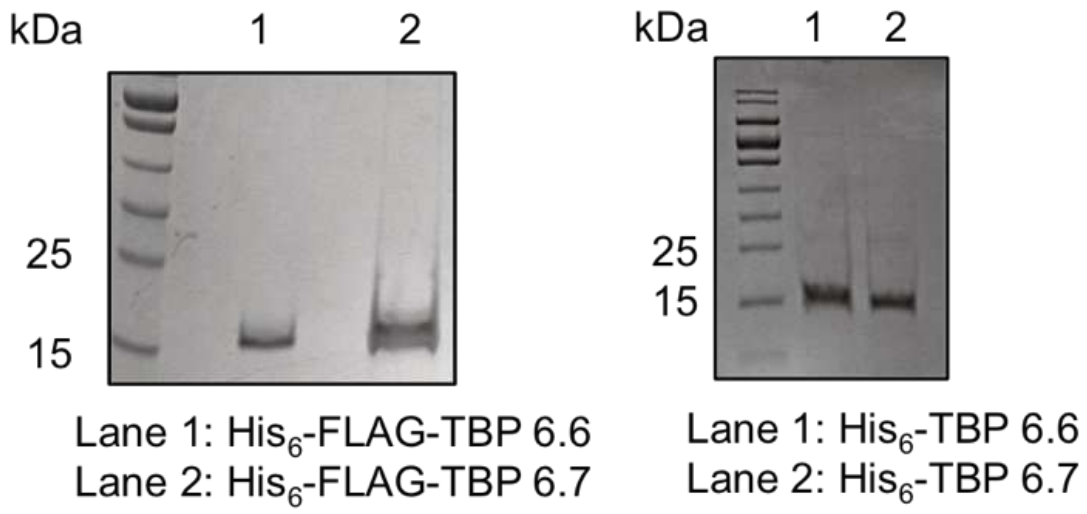


Figure 3.1: PAGE Gel of TBP 6.6 and TBP 6.7 This gel shows that TBP 6.6 and TBP 6.7 purify cleanly

Protein Purification Notes

The second dialysis is in phosphate buffer because initial ELISA experiments did not work in HEPES buffer. Since the screening was performed in PBS (obviously a phosphate buffer) and the phosphate buffer described had been working for other lab members' ELISA experiments, I decided to perform assays using this buffer, and began dialyzing my purified proteins against it to ease the transition.

Even though I needed the proteins in phosphate buffer for the ELISA assay, I continued to use the HEPES buffer for protein purifications. The general makeup of the buffer—Mg^{+2} ions and an EDTA chelator in a Good's buffer—is used to purify many RNA binding proteins [115–117]. The benefit of such a purification environment is less non-specific interaction between positively charged RRM and cellular nucleic acid, but the downside is that the presence of the chelator reduces yield in a nickel purification. This is a major downside but is necessary. I once attempted to purify TAR binding proteins using the normal phosphate buffer, but it resulted in the protein precipitating during dialysis.

3.3 ELISA Assay

The general strategy for my ELISA is shown in Figure 3.2. It is a fairly traditional sandwich ELISA, though it required a great deal of optimization.

Detailed assay notes follow, for anyone who may wish to build upon it.

3.3.1 General ELISA Protocol

Materials

The base for the buffer used in ELISA experiments was the second dialysis buffer (20 mM Phosphate, pH = 7.4, 150 mM NaCl) from the protein purification. The only practical consideration with doing this is that it is unwise to wash the proteins in a concentration column with this buffer if there is any possibility that a dialysis bag was leaking, since this will also concentrate the leaked protein into any samples. The ELISA buffer was 20 mM phosphate, 150 mM NaCl, pH = 7.4 with 0.05% Tween-20, and 0.1 mg/mL Bovine Serum Albumin (BSA). Generally, 250 mL of buffer was sufficient, and it was prepared by adding 125 µL of Tween-20 using a small syringe. The BSA was prepared by dissolving 100 mg of BSA into 10 mL of phosphate buffer, using 2.5 mL of this stock.

The assay was most successful when performed using streptavidin coated plates with only 5 picomoles (Thermo Scientific cat. 15124) of immobilized biotin, rather than the more commonly used plates with 10 picomoles (Thermo Scientific cat. 15125).

The best 3,3′,5,5′-Tetramethylbenzidine (TMB), in my experience, is TMB One, from Promega.

Plate Equilibration

In a typical experiment, the wells to be used (generally 4 rows, for a total of 48 wells, is the maximum number of wells able to be efficiently utilized) were incubated for 5 minutes with 200 µL ELISA buffer, washed by aspirating the buffer with a Pasteur pipet with vacuum, and incubated and washed a second time.

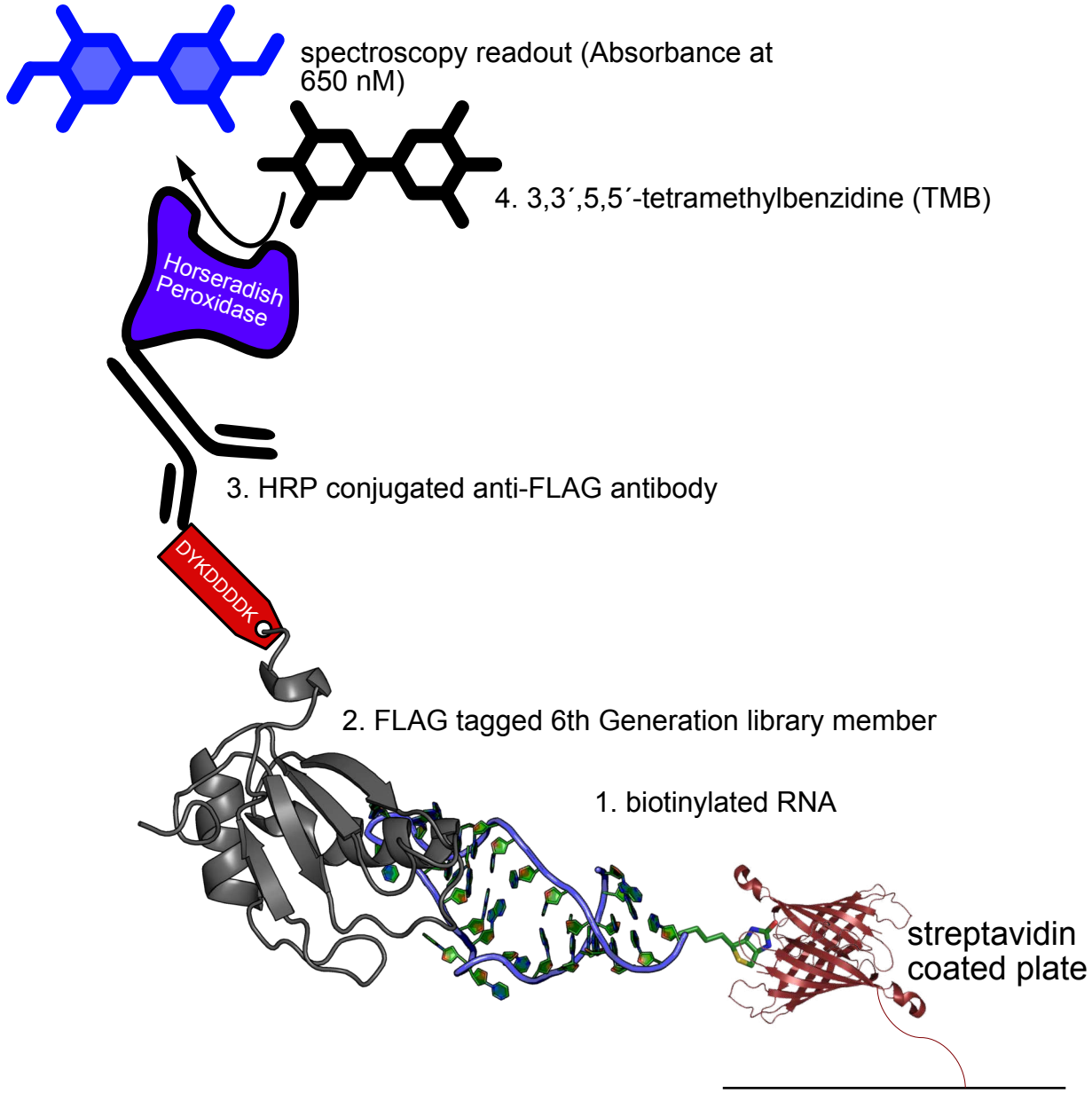


Figure 3.2: General ELISA Scheme An illustration of the scheme for the sandwich ELISA used to roughly, and eventually precisely, characterize the 6th generation TAR binding proteins

Generally speaking, "washing" refers to incubation for five minutes with 200 µL of ELISA buffer followed by aspiration. Buffer for washing was typically kept ice-cold, while "incubations" were at ambient room temperature. Plate was continuously shaking on an oval shaker at ~200 RPM.

The plate was then allowed to "equilibrate" to the buffer by incubating 200 µL of ELISA buffer for 1 hour. The buffer was aspirated, and the first incubation was begun.

RNA Preparation and Incubation

RNA was melted and refolded (2.4.2). All incubations were performed at 100 µL and 5–10 pmoles of RNA was required to saturate the plate, so the final 1X concentration of RNA needed was $1.1 \times \frac{5 \times 10^{-12} moles}{100 \times 10^{-9} L}$ = 55 nM if using the 5 pmol plates, or 110 nM if using 10 pmol plates. Given a stock concentration of 100 µM, this meant that the final dilution was ~1000-fold. Importantly, the first dilution, which brought the concentration to 100X (5.5–11 µM), should be done into buffer without Tween-20 and BSA, and the subsequent 100X dilution done into ELISA buffer shorty before application to plate.

100 µL of RNA was added to the plate, and allowed to incubate at room temperature for 2–4 hours. The RNA was aspirated from the plate using a Pasteur pipet attached to vacuum. The pipet was discarded and replaced between removal of different RNAs.

The plate was washed 3 times with ice-cold buffer prior to addition of protein.

Protein Preparation and Incubation

Protein concentrations were analyzed , and proteins were diluted to a 10X concentration in phosphate buffer not containing Tween-20 or BSA. These 10X protein solutions were distributed to strips of 300 µL PCR tubes which were arranged to have the same positioning as the ELISA plate. Using a 12-channel pipettor, *immediately* prior to incubation, ELISA buffer was added (9 times the volume to bring the 10X protein to 1X), dilution was mixed, and 100 µL added to the plates in the correct row. For example, if the assay was being performed with 50 nM protein, 25 µL of 500 nM protein would be added to the appropriate PCR tube, and 225 µL of ELISA buffer

added. The pipettor would then be adjusted to pipet 100 μL in each channel, and the solutions were applied to the plate.

Incubations were 1 hr. at room temperature, and were followed by 3 washes. If multiple concentrations of protein were used, aspiration began with the lowest concentration and continued to the highest. If switching between proteins or between RNAs, the pasteur pipet was washed by aspirating ~1 mL of 70% ethanol followed by ~1 mL of ddH_2O.

Antibody Incubation

The HRP conjugated anti-FLAG antibody was diluted in Odyssey Blocking Buffer (Li-Cor) at 1:10,000 (0.1 μL/mL), and mixed by rotation for ~10 minutes prior to application to plate. 100 μL of this mixture were applied to each well, and allowed to incubate at room temperature for 30 minutes. The antibody was removed, with pasteur pipet washed with ethanol and water between conditions, and the plate washed 4X.

TMB

TMB was allowed to come to room temperature for 30 minutes in a foil-wrapped 15 mL falcon tube. On the final antibody wash, wells were inspected for any residual liquid, and when dry, 100 μL of TMB added to each well. The plate was wrapped in foil, and allowed to incubate.

Absorbance was measured at 655 nM with a plate reader, and a higher absorbance is assumed to indicate a higher level of protein/RNA binding.

3.3.2 ELISA results

Initial attempts at ELISA were challenging, but a general protocol emerged using 5′ biotinylated TAR, or 5′ biotinylated U1hPII as the RNA, and incubating with a fixed concentration of various library members, or wtU1A. Initial "guesses" for a good concentration of protein, ~1 μM, were too high, and resulted in excessive background signal. The assay began to behave dynamically, and give good negatives when a concentration of 50 nM of protein was used. The results of the first such assay are shown in Figure 3.3.

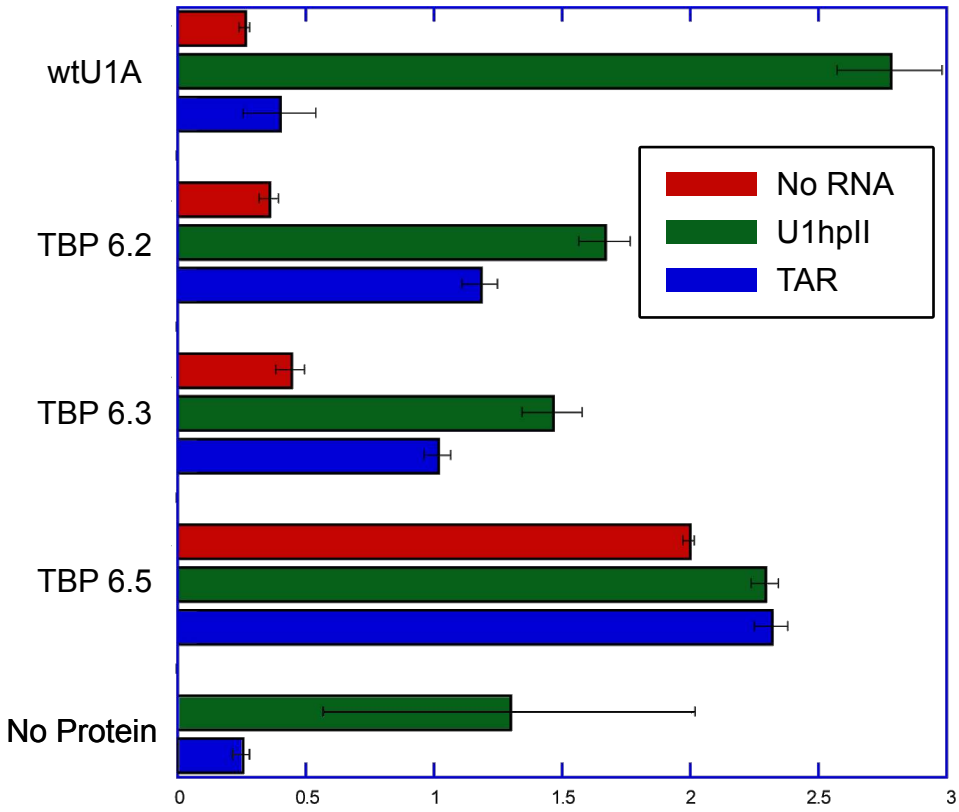


Figure 3.3: Results of a Single Plate of ELISA Assays using 50 nM U1A Variants The data shown in this assay represent the first time that wtU1A–U1hpII was used, and the high signal by that pair demonstrated that the assay was most likely reflecting reality. This validated the use of this assay and led to confidence in using it to compare the TBP 6th generation mutants, though none of the other mutants on this plate were particularly avid binders of TAR.

Though there was still non-specific background, there was also variation in signal. This concentration-dependent variation indicated that I was in or near the dynamic range of the assay. Additionallyk, the fact that the remarkably tight wtU1A–U1hpII interaction was generating high signal indicating that the assay was reflecting reality.

This assay, with 50 nM wtU1A or library member applied to immobilized TAR or U1hpII, was used to measure the relative TAR binding activity of library members, as well as to ensure that they did not bind U1hpII. Data was normalized to a positive control on each plate (wtU1A–U1hpII) in order to compare different assays. The results of these normalized ELISAs are shown in Figure 3.4. All but one library member (TBP 6.1) appreciably bound TAR at 50 M, and did not bind U1hpII significantly above background. These data are shown in Figure 3.4.

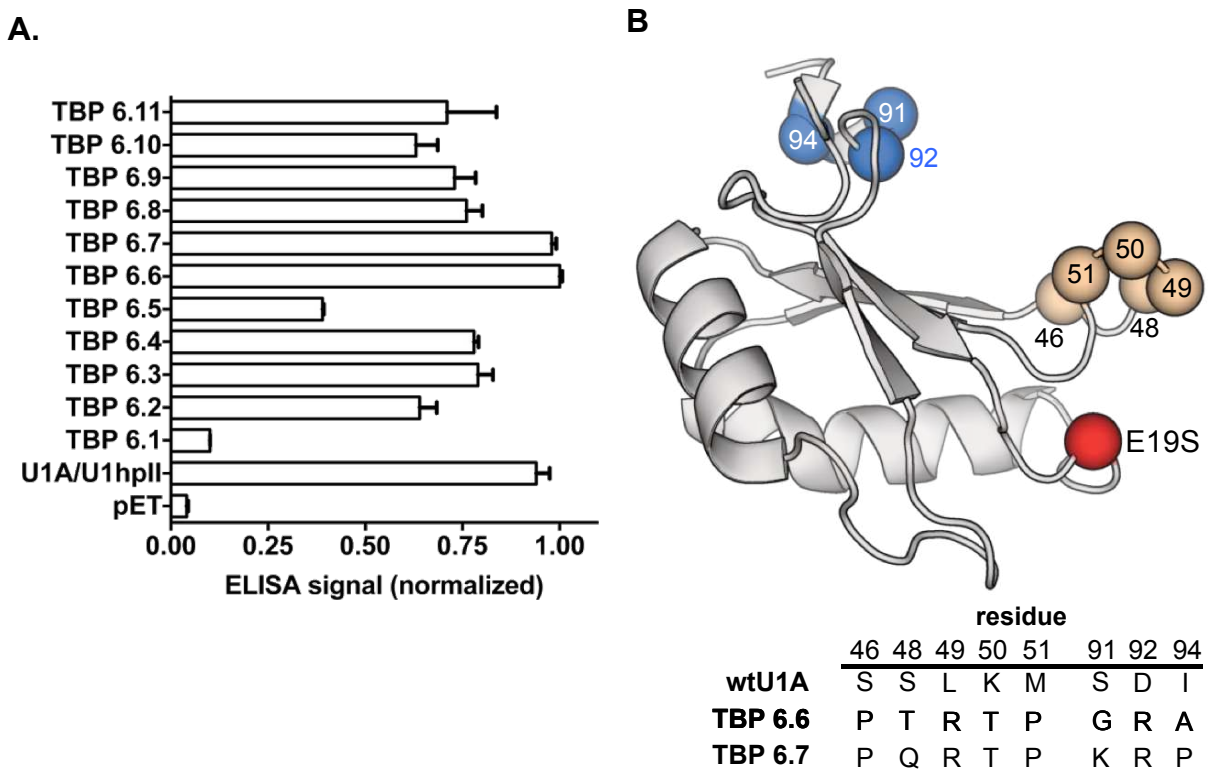


Figure 3.4: ELISA Survey of 6th Generation TAR Binding Proteins Shown in **A** is ELISA analysis of selected 6th generation TAR Binding Proteins. These data are the result of combining data from assays performed on different days, with the data from each day normalized to the wtU1A–U1hpII interaction. These data were then normalized to the highest normalized signal, that of TBP 6.7, which was given a value of "1." **B** Crystal Structure of U1A protein highlighting the mutations identified in the best apparent TBPs: TBP 6.6 and TBP 6.7 (adapted from [109])

Additionally, *E. coli* containing only empty pET plasmid were taken through the protein purification protocol as outlined in Section 3.2.3, concentrated to a normal degree (generally ~10-fold), and used in the ELISA experiment to ensure that our signal was not coming from non-specific interactions with trace *E. coli* proteins.

As a final qualitative check to be absolutely sure that the excellent results we were seeing were the result of the yeast display and selection, the most avid binder (TBP 6.7) would be compared to this pET "purification," and library foundation U1A E19S. As is clear from Figure 3.5, the *fact* of our success was clear, but we still needed to quantify it.

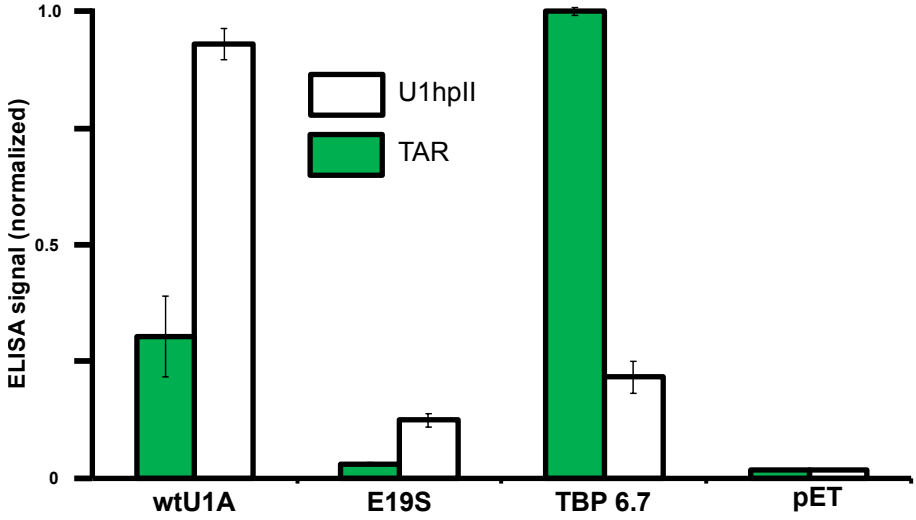


Figure 3.5: ELISA Signal of TBP 6.7 vs. U1A E19S for TAR Binding The most straightforward way to demonstrate the degree of improvement from the output of a library screen is to compare the scaffold and end-product head-to-head. We can see a remarkable change in TAR affinity from U1A E19S to TBP 6.7

Two particularly avid binders—TBP 6.6 and (especially) the aforementioned TBP 6.7—would become the basis for extensive future study.

3.3.3 Quantitative ELISAs

Though the ELISA assay seemed robust in comparing proteins to each other, it was initially unable to give us any absolute idea of the K_D value range for our TBP 6.X proteins. That would change while I was in the process of evaluating the approximate dynamic range of the ELISA

assay for the best binder to date (TBP 6.6). In so doing, it appeared that there was a hint of a binding curve present (seen in Figure 3.6A), with a desaturation and saturation point. I repeated the assay the next day, and used more concentrations in the apparent transition from unsaturated to saturated signal (corresponding to single digit nM). To my delight, when I plotted these new data a clear sigmoidal binding curve emerged (Figure 3.6b). Fitting with the Hill equation gave a K_D of ~5 nM.

I repeated the experiment for TBP 6.6 and, also for TBP 6.7, with all concentrations in triplicate. This confirmed the results for TBP 6.6, and established TBP 6.7 as having comparable affinity. These curves can be seen in Figure 3.7, and established final dissociation constants, by our analysis, of K_D = ~6 nM for TBP 6.6 and K_D = ~7 nM for TBP 6.7.

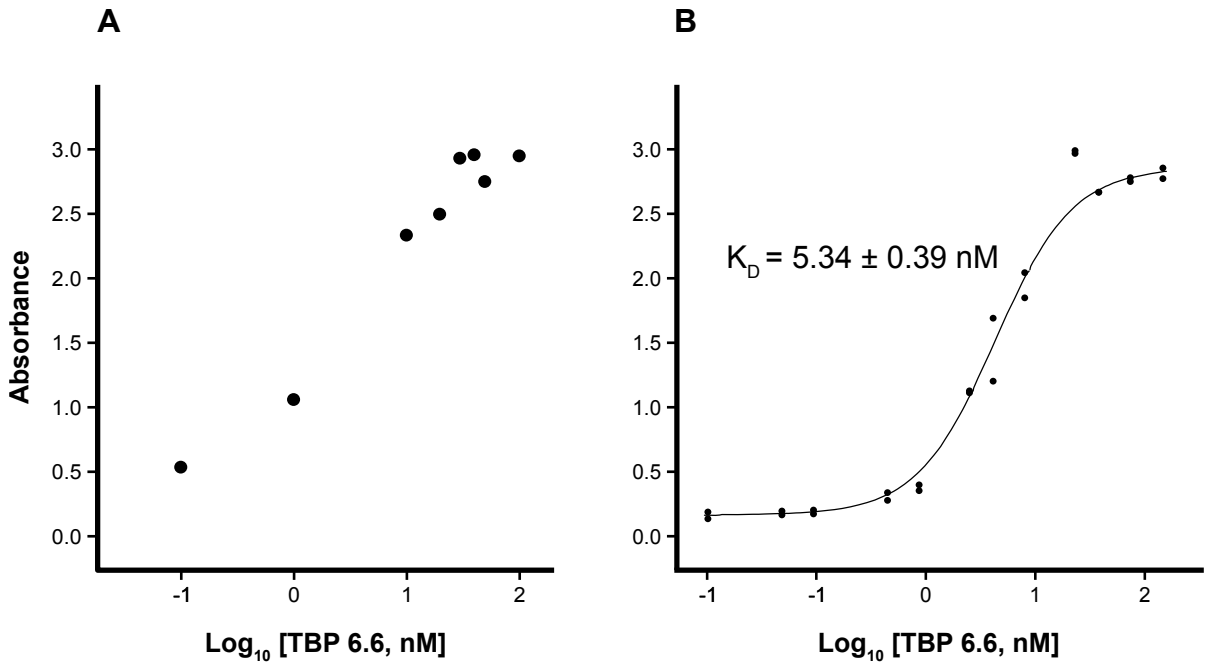


Figure 3.6: Initial Binding Curve Generated by ELISA A A semi-quantitative ELISA of TBP 6.6 showed promise, so I chose to **B** try again with more data points, and found that there was a believable binding curve

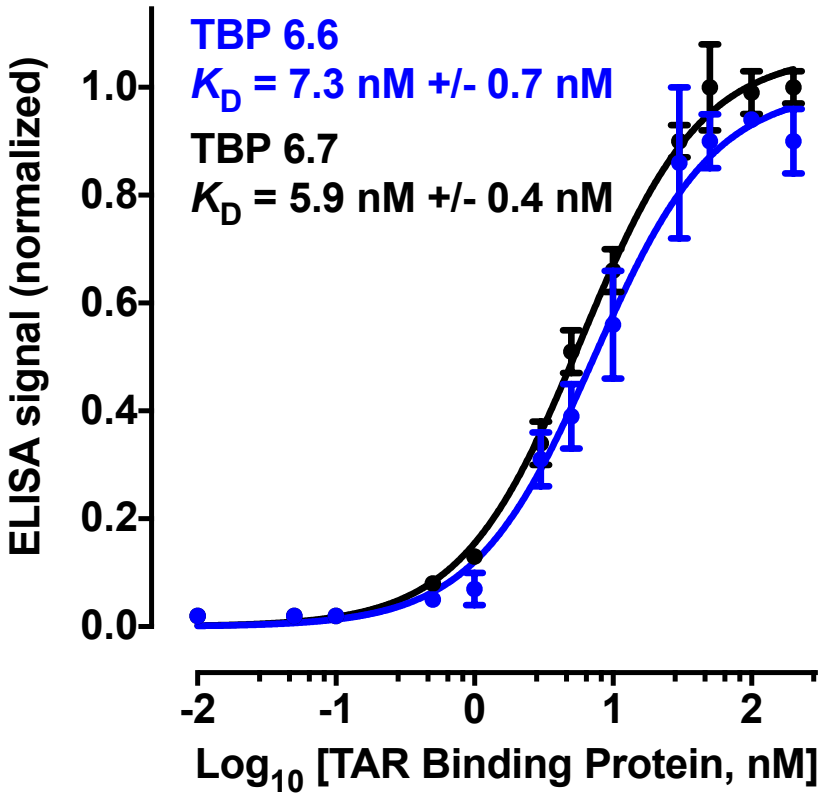


Figure 3.7: Finalized Quantitative ELISA-based Binding Curves The ELISA assay behaved *quantitatively*, unusual for a qualitative assay. When fit to the Hill Equation, the data seemed to indicate binding with K_D =1–10 nM. As a personal note, this is the data which I am most proud of obtaining. Adapted from [109].

3.4 SPR Analysis

3.4.1 Protein Preparation

With the repeatable data indicating that TBP 6.6 and TBP 6.7 were binding TAR with a dissociation value in the single-digit nM range. These values would represent the tightest binders of TAR RNA to date. In order to corroborate this extraordinary finding, I sent the proteins to the lab of Ite Laird-Offringa for analysis via Surface Plasmon Resonance (SPR).

The TBP 6.6 was sent for SPR analysis as the same C-terminal His-FLAG fusion protein used in ELISA analysis (Section 3.2.2). TBP 6.7 was, for final analysis, analyzed without the FLAG tag. This variant was prepared by PCR amplifying the TBP 6.7 C-term His-FLAG construct with the U1A-NcoI FP (5′-ATA TAC CAT GGC CCA GGT GCA GC-3′) and TBP 6.7-His RP (5′-GTT AAT TAA CTA TTA GTG ATG ATG GTG ATG ATG TGC GGC CGC AAC C-3′), cut with NcoI and PacI, ligated into pET plasmid, and purified as described in Section 3.2.3.

It is worth noting that concentrations of TBP 6.6 and TBP 6.7 were determined via A_{280} using a Nanodrop prior to shipping proteins. Initial SPR experiments were performed using this concentration as a given. A second experiment was performed on the TBP 6.7 in which concentration was determined by gel at the Laird-Offringa lab at USC, after shipment of protein. Due to differences in measurement and instability during shipment, this resulted in an ~2-fold lower concentration than was determined prior to shipment. In Table 3.2 this is referred to as "TBP 6.7 (Corrected Concentration)."

3.4.2 SPR Experiment

SPR was performed by Po Han Chen, in the Laird-Offringa lab at the University of Southern California. Though I was not involved in the performance of the assays, I selected and purified the proteins involved, and the results both corroborated and informed my own research.

A streptavidin-coated sensor chip (Sensor chip SA, GE Healthcare) was used to coat 25 response units (RU) of 5′-biotinylated U1hpII on flow cell 1, and 25 RU of TAR on flow cell 3, leaving flow cells 2 and 4 blank for background correction. Proteins were serially diluted in running

buffer (10 mM Tris-HCl, pH = 8, 150 mM NaCl, 5% glycerol, 62.6 μg/mL bovine serum albumin, 1 mM dithiothreitol, 0.05% surfactant P20, and 125 μg/mL yeast tRNA) to the concentrations indicated (Fig. 4D) and injected at 20 °C with a flow rate of 50 uL/ min over all surfaces consecutively. In each of three independent experiments, triplicate injections were fully randomized and interspersed with buffer injections to allow double referencing. After each protein injection the surface was regenerated with a 1-min 2 M NaCl injection, followed by a buffer injection. Data was processed using Scrubber and analyzed using CLAMP XP2 and a simple 1:1 Langmuir interaction model with a correction for mass transport. Association and dissociation rates are listed in Table S2. U1A RRM1 was injected for comparison, using the conditions described above. The biotinylated RNA was from the same stock as that used in the ELISA, demonstrating that RNA was viable.

3.4.3 SPR Analysis

A benefit of SPR is that it is able to give full kinetic information. A complete analysis with the k_a, the k_d, and the K_D for TBP 6.6–TAR, TBP 6.7–TAR, and U1A–U1hpII is shown in. Full statistical information for the initial, uncorrected experimental values, can be found in Table 3.1. The interesting contrast between TBP 6.6 and TBP 6.7 can only be seen in this kinetic data. Their K_D values are statistically identical, but their kinetic profiles are drastically different.

$$K_D = \frac{kd}{ka}$$

So a binding interaction is, intuitively, tighter if the complex forms faster, or dissociates slower. For the TBP 6.6–TAR complex:

$$K_D = \frac{1.56\times10^{-2}M^{-1}s^{-1}}{1.27\times10^{7}s^{-1}} = 1.3\,nM$$

While for the TBP 6.7–TAR complex:

$$K_D = \frac{1.56\times10^{-}2M^{-}1s^{-}1}{1.27\times10^{7}s^{-}1} = 1.5\,nM$$

Table 3.1: Statistical Values of SPR for TBP 6.6 and TBP 6.7 and wtU1A for TAR and U1hpII RNAs

TBP 6.6 and TAR RNA			
	experiment 1	experiment 2	experiment 3
k_a	1.16×10^{7}	1.23×10^{7}	1.42×10^{7}
k_d	1.91×10^{-2}	1.38×10^{-2}	1.39×10^{-2}
K_D (M)	1.65×10^{-9}	1.12×10^{-9}	9.79×10^{-10}
	average	SEM	SD
k_a	1.27×10^{7}	7.58×10^{5}	1.31×10^{6}
k_d	1.56×10^{-2}	1.76×10^{-3}	3.04×10^{-3}
K_D (M)	1.25×10^{-9}	2.03×10^{-10}	3.51489×10^{-10}
TBP 6.7 and TAR RNA			
	experiment 1	experiment 2	experiment 3
k_a	1.81×10^{7}	4.64×10^{7}	2.32×10^{7}
k_d	9.01×10^{-3}	2.31×10^{-2}	1.30×10^{-2}
K_D (M)	4.98×10^{-10}	4.98×10^{-10}	5.60×10^{-10}
	average	SEM	SD
k_a	2.92×10^{7}	8.72×10^{6}	1.51×10^{7}
k_d	1.50×10^{-2}	4.20×10^{-3}	7.28×10^{-3}
K_D (M)	5.19×10^{-10}	2.06×10^{-11}	3.55991×10^{-11}
wtU1A and U1hpII			
	experiment 1	experiment 2	experiment 3
k_a	1.21×10^{7}	7.13×10^{6}	1.38×10^{7}
k_d	5.41×10^{-4}	3.60×10^{-4}	3.64×10^{-4}
K_D (M)	4.47×10^{-11}	5.05×10^{-11}	2.64×10^{-11}
	average	SEM	SD
k_a	1.10×10^{7}	2.00×10^{6}	3.47×10^{6}
k_d	4.22×10^{-4}	5.97×10^{-5}	1.03×10^{-4}
K_D (M)	4.05×10^{-11}	7.27×10^{-12}	1.25876×10^{-11}

According to these data, which are the best head-to-head comparison between the two, the TBP 6.7 is ~5-fold slower to form, but also ~5-fold slower to dissociate. Notably, given equal K_Ds, slower kinetics are preferable in a pharmacological sense, since a slower on/off rate means that an effector molecule *stays* with its target. Additionally, slower kinetics generally indicate better pharmacological specificity [118]. In a more general sense, the fact that the kinetics vary between TBP 6.6 and TBP 6.7 indicates that they have slightly different modes of binding.

Figure Analysis

The SPR binding curves can be seen in Figure 3.8 and demonstrate the excellent binding of TBP 6.6 and 6.7 to TAR RNA.

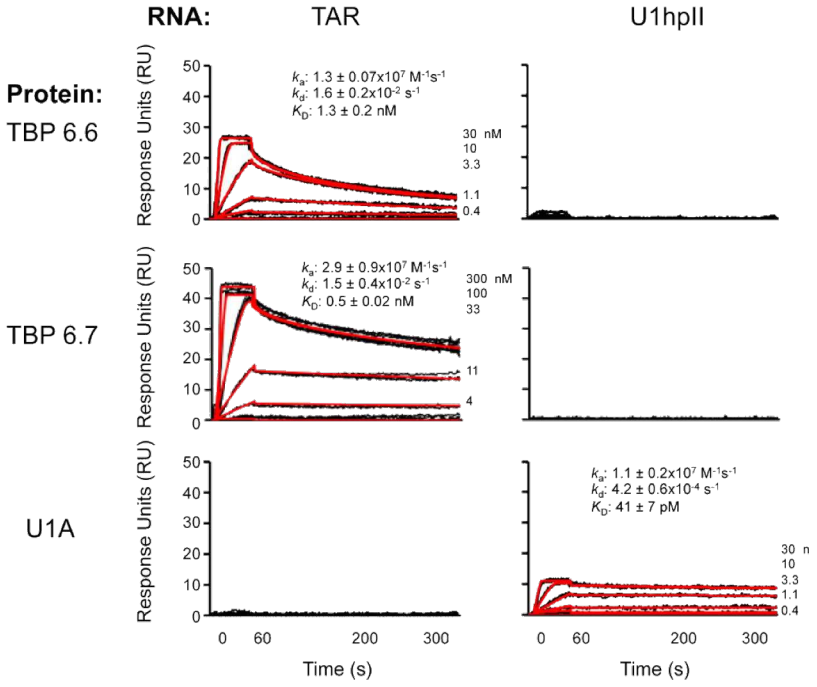


Figure 3.8: SPR Binding Curves of TBP 6.6, TBP 6.7 and wtU1A against TAR and U1hpII The binding and un-binding curves from the SPR experiment performed in the Laird-Offringa lab (adapted from [109]

The interaction between TBP 6.6 or TBP 6.7 and TAR mirrors the excellent affinity of U1A for U1hpII, while both the TBPs and U1A show little to no affinity for the off-target surface.

The final measurement of TBP 6.7–TAR binding, done with a corrected concentration to bring the values in line with normal Laird-Offringa lab concentration protocols, found that the interaction between TBP 6.7 and TAR was not low nM, but ~500 pM, a success beyond our wildest expectations. These data can be found in Table 3.2.

Table 3.2: Kinetic and Equilibrium Values of SPR for TBP 6.6, TBP 6.7, and wtU1A for TAR RNA and U1hpII RNA

	TBP 6.6–TAR	TBP 6.7–TAR	TBP 6.7–TAR (Adjusted Concentration)	U1A–U1hpII
K_D	1.3 ± 0.2 nM	1.5 ± 0.3 nM	0.50 ± 0.02 nM	0.041 ± 0.007 nM
k_a	$1.3 \pm 0.1 \times 10^{7}$	$5.6 \pm 0.7 \times 10^{6}$	$2.9 \pm 0.9 \times 10^{7}$	$1.1 \pm 0.2 \times 10^{7}$
k_d	$1.6 \pm 0.2 \times 10^{-2}$	$8.0 \pm 0.9 \times 10^{-3}$	$1.5 \pm 0.4 \times 10^{-2}$	$4.2 \pm 0.6 \times 10^{-4}$

3.5 Characterization of TAR Binding Selectivity

Binding selectivity, and the requirements for TAR RNA recognition, was further characterized by ELISA using a variety of TAR derived RNAs. The protocol was as described in Section 3.3.1, but with a standard protein concentration of 20 nM, rather the the 50 nM used in initial characterization. We used a TAR-derived RNA hairpin, designated hairpin 1 (hp1) that lacks the UCU bulge (Figure 3.9), and three TAR-derived RNAs designated hairpins 2, 3, and 4 (hp2, hp3, and hp4), which have two consecutive mutations in the apical loop (structures shown in Figure 3.9A, with mFold analyses shown in Figure 3.10 and Table 3.3.

The most dramatic change in binding was observed when we removed the UCU bulge in TAR (hp1). When we performed an ELISA using this immobilized RNA, essentially no binding was observed with either TBP 6.6 or TBP 6.7 (Figure 3.9,). This finding is important because the native TAR-binding protein (Tat) largely recognizes the UCU bulge. Thus, if our proteins occupy this space, they should antagonize the Tat–TAR interaction.

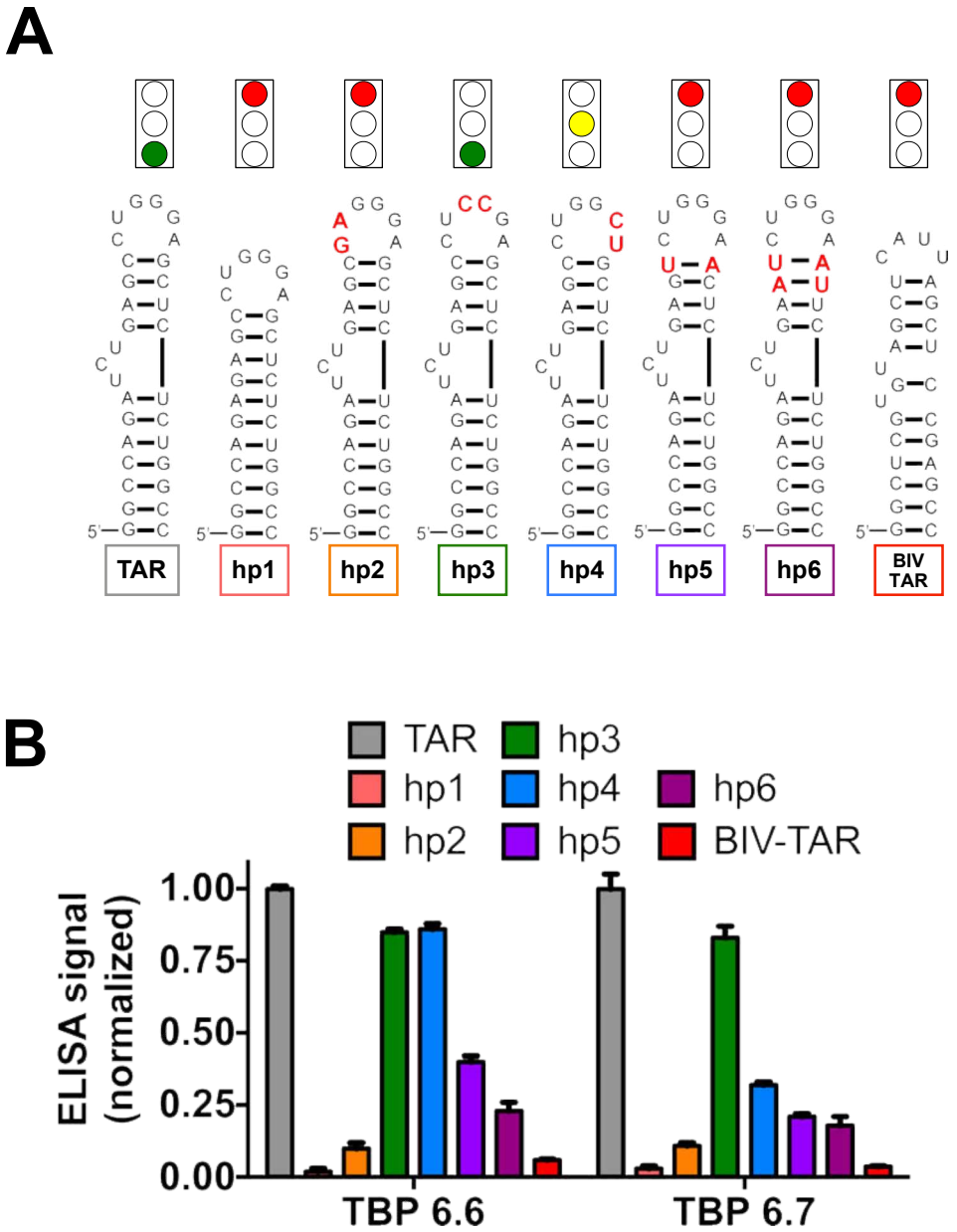


Figure 3.9: Affinity of TBP 6.6 and TBP 6.7 for Modified TAR TBP 6.6 and 6.7 show different affinities, via ELISA assays, for TAR derived hairpins with a variety of modifications, shown in **A** made to the stem-bulge, loop, or stem with the intention of probing binding mechanism. We also tested the affinity of TBP 6.6 using a different subtype of TAR, from the Bovine Immunodeficiency Virus (BIV-TAR).**B** shows the reduction in binding from these various changes. Of note is hp4, a loop mutation which greatly reduces binding to TBP 6.7, but not TBP 6.6 (adapted from [109]

Table 3.3: Folding Energies of RNA Hairpins Used in Selective Binding Studies

RNA	ΔG (kcal/mol)
TAR	-13.1
HP1	-18.9
HP2	-14.3
HP3	-13.1
HP4	-12.9
HP5	-11.1
HP6	-8.9
BIV-TAR	-12.2

Dramatically decreased binding was also observed when the first two loop nucleotides (5′-CU-3′) were mutated to 5′-GA-3′ (hp2). Less dramatic, but significant, changes in affinity were observed when we mutated other loop residues. When the last two loop nucleotides (5′-GA-3′) are mutated to 5′-CU-3′ in hp4, although no appreciable loss in affinity was observed for TBP 6.6, significantly lower binding was observed for the complex involving this RNA and TBP 6.7 (Figure 3.9B, blue bar). In contrast, mutating the central nucleotides (5′-GG-3′) to 5′-CC-3′ did not appreciably alter affinity for either protein (Figure 3.9B, green bar).

The ability of synthetic RRMs to recognize specific loop nucleotides begs the question: are nucleotides in the top of the stem (which link the bulge and loop) recognized by U1A-derived proteins? To probe this, we assessed binding of mutants hp5 and hp6, which have mutated residues in the top of the stem. Both mutations significantly decreased affinity to TBP 6.6 and TBP 6.7, suggesting that the uppermost stem nucleotides directly participate in complex formation. Additionally, we also tested binding to the TAR sequence from the bovine immunodeficiency virus (BIV TAR)(53). This RNA is structurally similar to HIV TAR but differs in the sequence and size of the loop (4 bases in BIV TAR, versus 6 bases in HIV-1 TAR) and in the nature of the stem bulge. We found that neither TBP 6.6 nor TBP 6.7 had appreciable affinity for this RNA .

This allowed us to make certain guesses about the nature of the interaction between TAR and TBP 6.6 or TBP 6.7, though they would remain guesses for a while longer. It *was* clear, based on the differences in binding to hp4 and the different kinetic profiles, that the two high-affinity

proteins had different modes of binding, and that the binding was sensitive to alterations in TAR, and therefore specific.

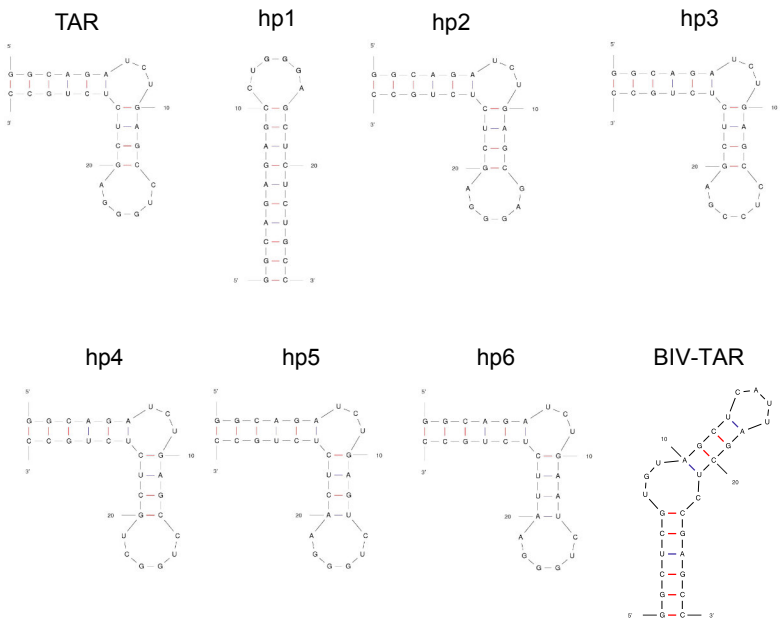


Figure 3.10: mFold Calculations of Hairpins Used to Analyze Selectivity We wanted to be sure that our tests were still being done with folded RNA, and mFold calulations indicate that given our melting and refolding protocol, there was little danger of the RNA being unfolded. Folding energies are given in Table 3.3

3.6 SHAPE Analysis

I was also fortunate to have the assistance of Chringma Sherpa and Stuart Le Grice at the National Cancer Institute, who were able to perform selective 2′hydroxyl acylation analyzed by primer extension (SHAPE) analysis on the interaction of TBP 6.6 and 6.7 with the HIV 5′ UTR. SHAPE is a complicated assay, but can be generally described as using the accessibility of the 2′-OH on RNA as a proxy for conformational flexibility at a given base, and changes in conformational flexibility upon binding with protein are used as an indicator of where the protein binds.

3.6.1 SHAPE Context

We predicted that TBP binding to TAR would make the TAR region on the HIV 5′ UTR less amenable to chemical modification. We studied binding in the context of the 362-nt 5′-UTR because this RNA is highly structured, harboring the TAR and polyA hairpins, the primer binding site (PBS), packaging signal (Ψ), dimer linkage sequence (DLS), and major 5′ splice site (5′ss). [119, 120]. These multiple cis-acting elements bind different ligands and support long-range interand intramolecular interaction(s) that facilitate genome transcription, translation, RNA dimerization/packaging, and splicing. Therefore, TBP binding to the 5′-UTR provides a direct and biologically relevant measure of potential off-target effects.

3.6.2 SHAPE RNA Prep

The 5′-UTR RNA exists in monomeric and dimeric forms [121–124], and because SHAPE measures ensemble-average reactivity, folding conditions were optimized to prepare a homogeneous monomeric RNA Figure 3.11A.

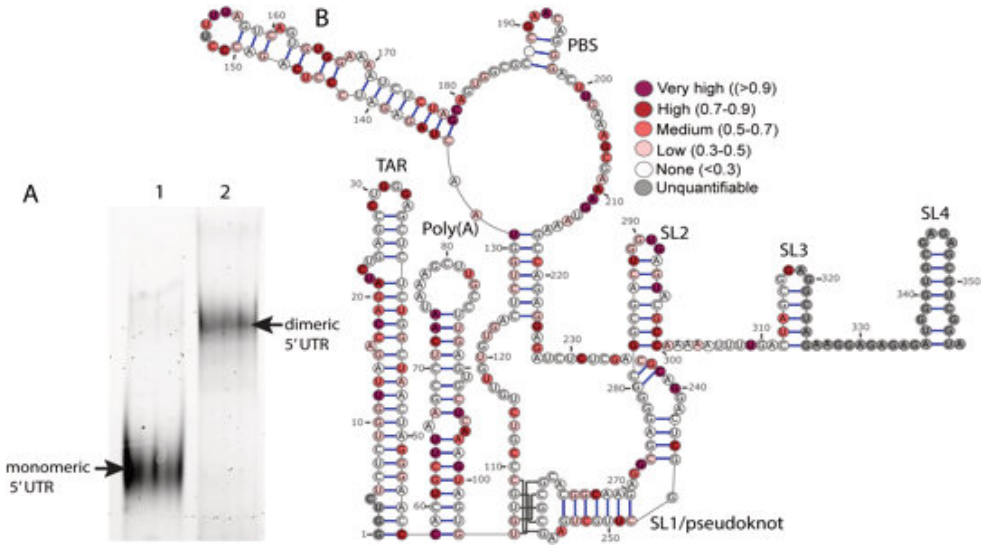


Figure 3.11: RNA Folding Conditions and Baseline Reactivities for SHAPE In **A** we an see the clear difference in size on a PAGE gel of monomeric vs. dimeric HIV 5′ UTR, and in **B** the baseline SHAPE reactivities of the HIV 5′ UTR. Adapted from [109].

To prepare monomeric RNA, the 362 nt long NL4-3 5′ UTR RNA was prepared by in vitro transcription using the MegaShortScript kit (Ambion/Life Technologies) according to manufacturers' recommendations. DNA template used in the transcription reaction was generated by PCR from a proviral pNL4-3 plasmid using high fidelity platinum Taq DNA polymerase (Invitrogen) and forward and reverse primers "T7L" (5′-TAATACGACTCACTATAGGTCTCTCTG-3′) and "369R" (5′-GCTTAATACCGACGCTCTCGC-3′) respectively. The forward primer was designed to introduce T7 promoter sequence at the 5′ end of the 5′ UTR. The RNA was then treated with Turbo DNase I for 1 hour at 37 °C, heated at 85 °C for 2 min and run on a denaturing gel (5% polyacrylamide (19:1)), 1x TBE, 7M urea) at constant temperature (45°C, 30W max). The 5′ UTR band was then excised, electro-eluted at 200 V for 2 hours at 4 °C, ethanol precipitated and stored at -20 °C in TE buffer (10 mM Tris, pH = 7.6; 0.1 mM EDTA) prior to use.

3.6.3 SHAPE Method

In 5 different tubes, 40 pmoles of RNA in a total volume of 10 µl was refolded by heating to 85 °C for 2 minutes, followed by slow cooling to 25 °C for 15 minutes (ramp rate 0.1°C/sec). Meanwhile, two-fold serial dilutions of U1A mutant protein (20 picomoles/µl, 10 picomoles /µl , 5 picomoles /µl , 2.5 picomoles /µl) were made in the protein storage buffer (20 mM phosphate, pH = 7.4, 150 mM NaCl containing 10% glycerol). The volume in each tube was brought to 284 µl by adding 274 µl of nuclease free water (Invitrogen). 16 µl of each protein dilution or 16 µl of protein storage buffer alone was incubated with the folded RNA at 37 °C for 10 mins. 144 µl of each RNA-protein mixture was aliquoted into two tubes labeled as "NMIA+" and "NMIA-". RNA in the "NMIA+" tubes was chemically modified by incubation with 16 µl of 30 nM NMIA in anhydrous DMSO at 37 °C for 20 minutes. To the "NMIA-" tubes, 16 µl of anhydrous DMSO was added and these tubes were also incubated at 37 °C for 20 minutes. Protein in both the "NMIA+" and "NMIA-" tubes was removed by phenol-chloroform extraction. For this, 140 µl of water followed by 300 µl of phenol:chloroform:iso-amyl alcohol mixture pH 6.8 (Ambion) was added to each tube and spun at 14000 rpm at 4 °C for 5 mins. 250 µl of the aqueous phase was recovered

and ethanol precipitated. The RNA pellet was suspended in 12 µl nuclease free water. 3 picomoles of each RNA were then used to generate cDNA library for each RNA. Subsequent cDNA processing/fractionation and SHAPE data analysis were conducted as previously described [125].

To check the homogeneity of the RNA samples, 20 µl of the SHAPE reaction mix containing 16 µl of protein storage buffer alone was sampled out just before the addition of NMIA and fractionated on a native gel [4% polyacrylamide (19:1), 1x TBE] at constant voltage of 200V at 4 °C for 5.5 hours. RNA bands were visualized by SYBR Green II RNA Gel Stain (Life Technologies).

3.6.4 SHAPE Results

Averaged SHAPE reactivity values from three independent experiments were color-coded onto the proposed pseudoknot monomeric 5′-UTR structure [122,123] as the algorithm of the software commonly used for RNA secondary structure prediction (RNAStructure) cannot be used to predict pseudoknot structures. As shown in Figure 3.11B, data for the 5′-UTR RNA in the absence of protein are at slight variance with the proposed pseudoknot structure.

Such discrepancy in HIV-1 monomeric RNA secondary structure, which was previously reported [123], may reflect differences in folding conditions and tertiary interactions. SHAPE analysis was performed at different RNA:TBP ratios (1:1, 1:2, 1:4, and 1:8), and appreciable changes in acylation sensitivity were observed only when either protein was present in a 4- or 8-fold molar excess (Figure 3.12, Figure 3.14). Reactivity values were color-coded (Figure 3.14A for TBP 6.6 and Figure 3.14 for TBP 6.7) and plotted as a function of nucleotide position (Figure 3.14B for TBP 6.6 and Figure 3.14D for TBP 6.7). For the determination of which nucleotides were conformationally flexible or constrained by TBP binding, reactivity values in the absence of the protein were subtracted from those in its presence, and the resulting difference values were plotted as a function of nucleotide position (Figure 3.14C for TBP 6.6 and Figure 3.14F for TBP 6.7). Following the previously reported quantitative SHAPE difference cut-offs, 60 nucleotides with a reactivity difference > +1 were designated as conformationally more flexible and those with a reactivity difference < –1 were assigned as conformationally more constrained.

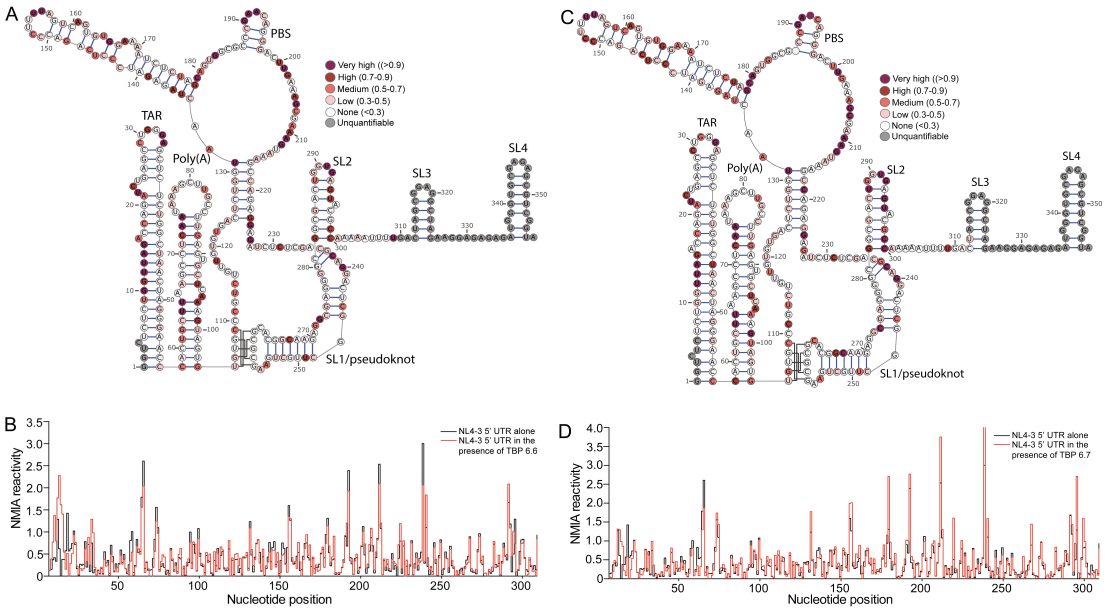


Figure 3.12: SHAPE Data Using 4:1 Protein:RNA A,B TBP 6.6–TAR or **C,D**TBP 6.7–TAR show some changes, but were less active than the 8:1 data shown in Figure 3.14. **B** and **D** show the reactivity of each base position, while **A** and **C** transpose this reactivity onto a map of the HIV 5′-UTR. Adapted from [109].

3.6.5 Qualitative Binding

One reviewer complaint about our affinity measurements is that they were all performed in the context a solid scaffold (both ELISA and SPR require immobilzied target). In the course of performing the SHAPE assay, qualitative gel-shift assays were performed which demonstrate binding outside the presence of a scaffolding. These data can be seen in Figure 3.13.

3.6.6 SHAPE Data

A common feature of TBP 6.6 and 6.7 complexes was increased acylation at several important positions of the TAR hairpin. Protein binding destabilized the local helix at nucleotides U 12 , U 13 , A 14 , G 15 for TBP 6.6 and U 13 , and A 14 for TBP 6.7. Nucleotide C 23 of the UCU loop (which was deleted in the hp1 mutant) was also rendered more flexible in the presence of TBP 6.6. Interestingly, of all TAR mutants tested for reduced binding to TBP 6.6 and 6.7 by ELISA (Figure 3.9), the hp1 mutant most significantly disrupts the interaction. Therefore, the SHAPE and ELISA assays collectively implicate UCU loop nucleotide C 23 in TBP 6.6 binding. TBP 6.7 significantly constrained nucleotide C18 located at the base stem implying C18 as an important

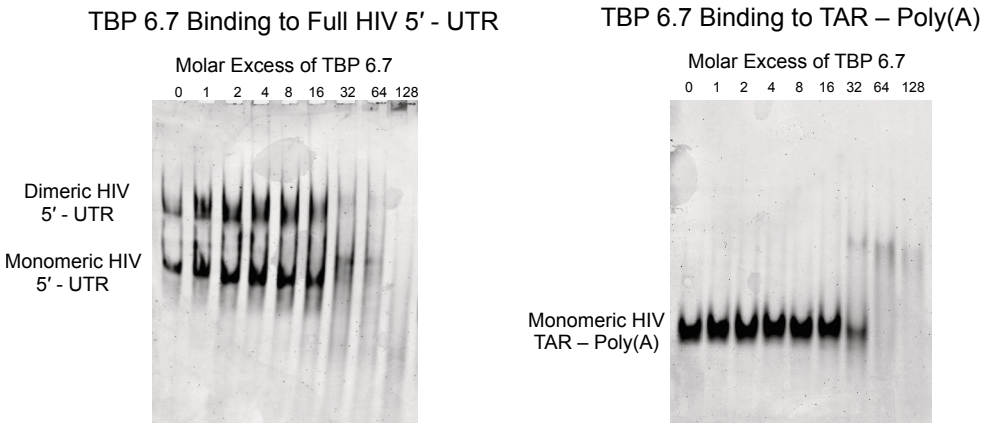


Figure 3.13: TBP 6.7 Binding TAR via Gel Shift Assay In both the full HIV 5′-UTR and the isolated TAR–Poly(A) element, appreciable shifts are seen with a 32-fold excess of TBP 6.7. This shift occurs in the full 5′-UTR in both the monomeric and dimeric forms. It is clear that this is not degradation, as all shifts appear to involve increasing mass. This qualitative data shows binding occurring in the absence of a solid support. Adapted from [109].

contact site for the protein. Conversely, no TAR nucleotide was rendered more constrained by TBP 6.6 binding, suggesting it may bind to an already constrained (base-paired) region, e.g., the upper stem, which was shown to be important for binding by ELISA (Figure 3.9). Therefore, both TBP 6.6 and TBP 6.7 proteins induce significant, yet distinctly different, conformational changes in TAR, suggesting they interact differently.

No significant changes in acylation sensitivity were observed outside the TAR region for TBP 6.6, indicative of a local interaction. In contrast, TBP 6.7 significantly increased conformational flexibility at nucleotides C58, A73, A74, U94, G99, U100, U131, C159, U176, G178, C179, A192, A211, G212, C219, C238, C267, and U295. TBP 6.7 also significantly constrained nucleotide C 292 in the SL2 loop (the major 5′-splice site). SL2 residues were shown to mediate long-range contact with residues at the base of SL1 and upstream of the U5 region 56 in the dimeric UTR. Therefore, by decreasing reactivity of the SL2 loop residue, TBP 6.7 could shift the equilibrium toward the dimeric UTR conformer.

Furthermore, we rule out nonspecific protein–RNA interactions driving the conformational flexibility observed outside the TAR hairpin of the 5′-UTR as TBP 6.7 strongly discriminates against (a) BIV TAR, a highly homologous relative of HIV-1 TAR (Figure 3.9) and (b) 5000-fold

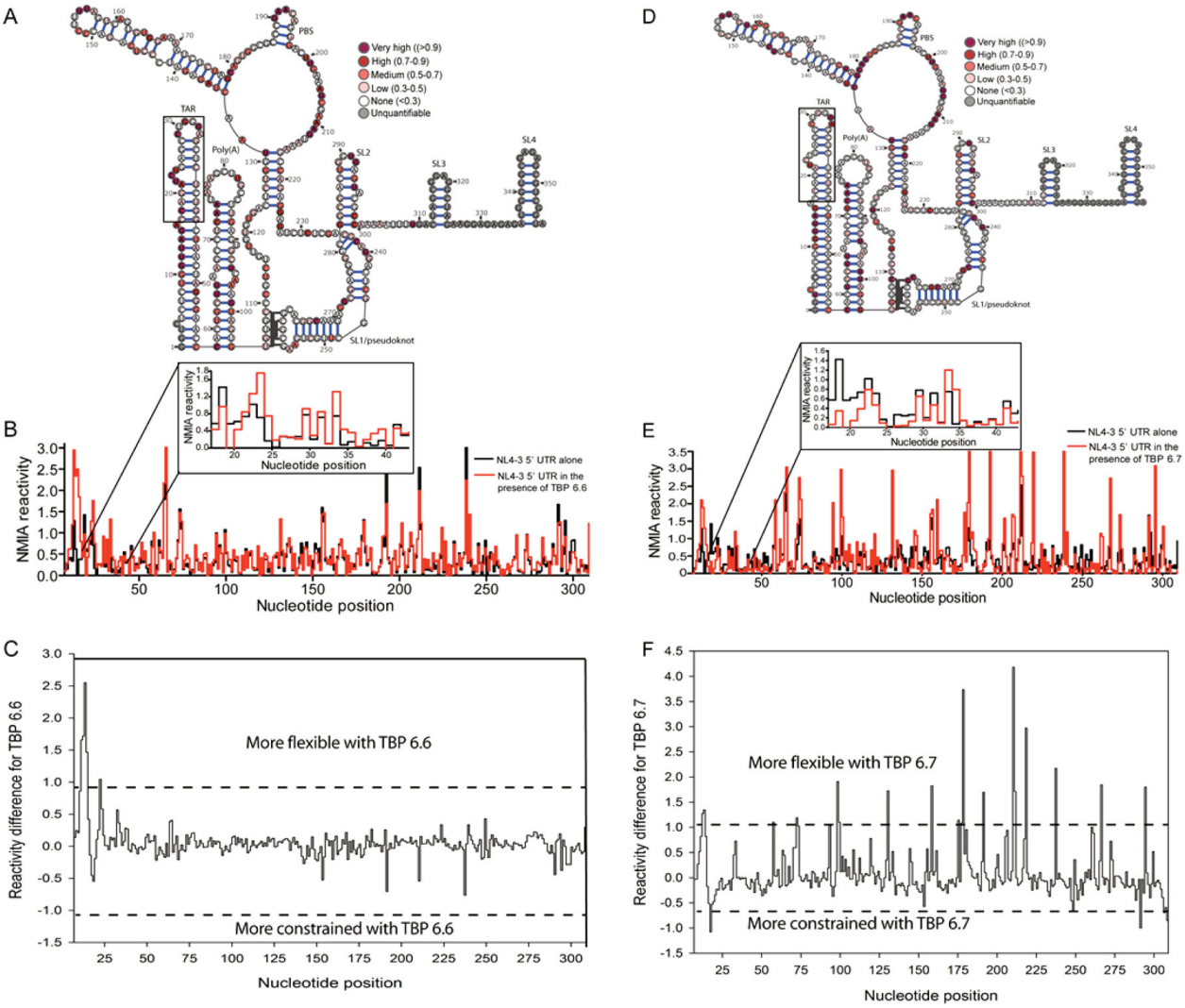


Figure 3.14: SHAPE Data Using 8:1 Protein:RNA With a protein:RNA ratio of 8:1 **A,B,C** TBP 6.6 or **D,E,F** TBP 6.7. The SHAPE profile of the 5′ UTR was significantly altered. The TAR regions in both were heavily impacted, but TBP 6.7 seemed to alter reactivity on residues far (in sequence space) from the TAR element. **B** and **E** show the reactivity of each base position in TAR, while **A** and **D** transpose this reactivity onto a map of the HIV 5′-UTR. **C** and **F** interpret these data as nucleotide positions which are more or less constrained. Adapted from [109].

excess of competitor tRNA (Figure 2.5). Therefore, we propose that TBP 6.7 binding to TAR in the context of the 5′-UTR induces long-range alterations in overall topology that are more pronounced than those promoted by TBP 6.6. Stated differently, secondary consequences of TBP 6.7 binding to TAR on global 5′-UTR topology cannot be ruled out. This notion is supported by recent work that identified small-molecule ligands specific for TAR, where changes in SHAPE reactivity profiles where observed distal to the ligand binding site [126].

Thus, we believe that differences in the nature of the primary interaction between TPB 6.6–TAR and TPB 6.7–TAR, and the structural consequences thereof, might explain why TPB 6.7 selectively induces long-range topological changes. These differences might also explain why, despite having equal affinity toward TAR RNA, the two TAR-binding proteins have different biological activity. As shown subsequently in the manuscript, TBP 6.7 prevents Tat–TAR interaction and inhibits transcription from the TAR region whereas TBP 6.6 does not (shown in Figure 3.15).

3.7 Disrupting the Tat–TAR Interaction

The synthetic proteins described in this work recognize TAR RNA with excellent affinity and exquisite selectivity. Though this is an achievement in-and-of itself, any potential therapeutic and many basic research utilities of these new proteins rests on their ability to inhibit a protein–RNA interaction involving the trans-activator of transcription (Tat) protein and TAR RNA. In binding to TAR, Tat alters the transcription complex, recruits the positive transcription elongation complex (PTEFb) of cellular CDK9 and cyclin T1, resulting in an increase in the production of full-length viral RNA [127]. Reagents that inhibit the Tat–TAR complex can suppress the transcription of full-length HIV RNA, leading to suppression or abrogation of HIV protein expression and production of virus [127].

To determine if our new TAR-binding proteins inhibit an interaction with Tat, we utilized a previously described Tat peptide comprising a portion of the full length Tat protein known to bind RNA (N-RKKRRQRRRRPPQSQTHQVSLSKQPTSQPRGDPTGPKE-C) [128].

3.7.1 ELISA

I initially attempted to use ELISA (as in Section 3.3) to determine if TBP 6.7 could disrupt or inhibit the Tat–TAR complex; however, the high theoretical charge of this Tat peptide N-RKK RRQRRRPPQGSQTHQVSLSKQPTSQPRGDPTGPKE-C; theoretical charge = +9) complicated these experiments. I found that Tat peptide adsorbed onto the plates whether or not biotinylated TAR was also immobilized (presumably through nonselective interactions with streptavidin) and could not be easily removed. Despite many attempted variants on the ELISA protocol, this simple fact of the Tat peptide *always* binding the plate meant that these attempts were not succesful.

3.7.2 ITC

To overcome the fact that our principal immobilized surface assay was ineffective, we used a solution phase experiment—isothermal titration calorimetry (ITC)—to characterize the interaction and the effect our TAR-binding proteins have on complex formation.

ITC was performed using a MicroCal iTC200 calorimeter maintained at 25 °C. TBP 6.6 and TBP 6.7 were expressed with C-terminal His tags, and purified by as described above (Section 3.2.3). Purified proteins were dialyzed overnight in phosphate buffer (20 mM sodium phosphate, pH = 7.4, 150 mM NaCl). Truncated HIV Tat peptide was ordered (Genscript Corp), and resuspended in this phosphate buffer. HIV TAR RNA was placed in the sample cell at concentrations ~6 μM, and ~60 μM Tat peptide was injected in 2.49 μL increments (16 injections total), with an initial injection of 0.4 μL, at 180 second intervals using a stirring speed of 750 rpm. Displacement experiments were performed by titrating 65.3 μM Tat peptide into a pre-formed 1:1 complex of TAR and TBP 6.7 (6.1 μM each), or a pre-formed 1:2 complex of TAR and TBP 6.6 (6.1 μM TAR and 13.8 μM TBP 6.6). Pre-formed complexes were used in order to reduce the complexity of the system to manageable levels: a displacement assay (with the TBP displacing Tat peptide) would have unpredictable heats of formation/dissolution, and mixing of the Tat and TBPs would be a complicated three-body system. Therefore, we pre-complexed the Tat and TBP in order to best

represent the most realistic cellular scenario where TBP 6.7 would *prevent* the Tat–TAR interaction rather than break it up.

Data were analyzed using Origin7.0 (MicroCal, ITC200) using a one set of sites binding model for fitting. All data were reference subtracted by subtracting the mean heat of dilution from each data point.

Figure 3.15 shows that TBP 6.7, but not TBP 6.6 block formation of the Tat–TAR complex. Presumably, this discrepancy relates to their apparently different modes of binding (given the differences in kinetics via SPR, and the differences in binding to TAR mutants). It could be as simple as the slower off-rate of the TBP 6.7 leaving the TAR free for TAT binding. It could also be that the different modes of binding by TBP may leave the mode of binding for TAT accessible for one, but not the other, or it could be that TBP 6.6 somehow recruits TAT peptide.

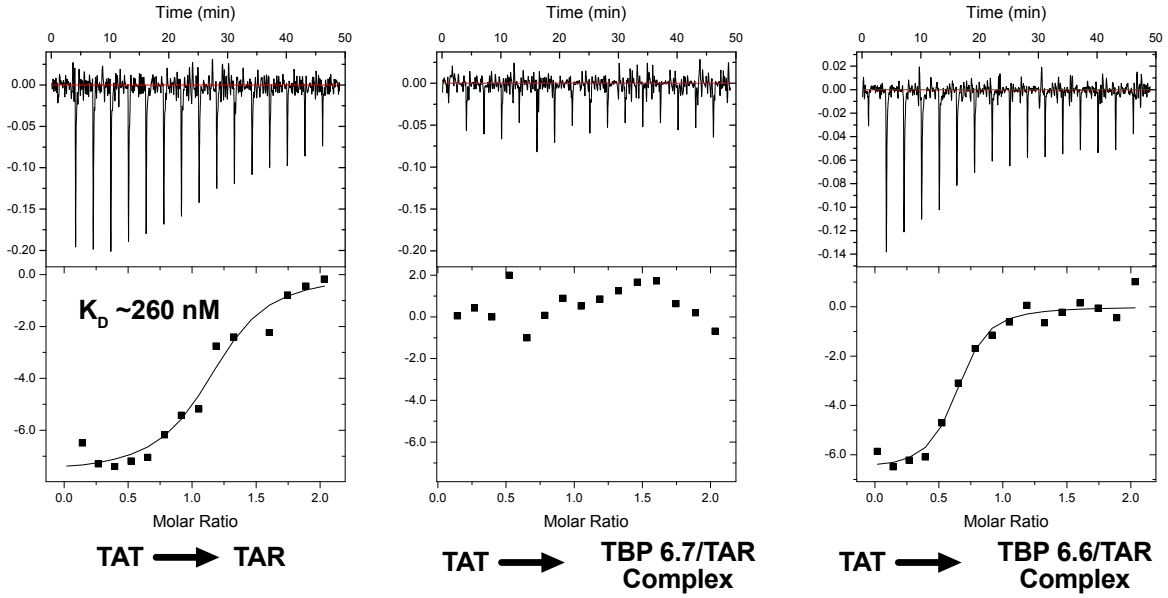


Figure 3.15: Disruption of Tat–TAR interaction Measured by ITC The RNA binding region of TAT peptide binds TAR RNA, but this binding does not occur if TAR is pre-complexed with TBP 6.7. Interestingly, pre-complexing with TBP 6.6 seems to have the opposite effect, possibly because the TBP 6.6 makes the TAR bulge more accessible, or because the TBP 6.6 recruits TAT peptide. Adapted from [109].

3.8 Suppression of Tat–TAR-Dependent Transcription by a Synthetic TAR-Binding Protein

I performed an *in vitro* transcription assay to quantify the suppression of Tat/TAR-dependent transcription of a portion of the HIV-1 genomic DNA which includes the TAR element [129–131]. I performed this *in vitro* transcription assay with HeLa cell nuclear extract in the presence of all the elements needed for Tat/TAR-dependent transcription, with and without TBP 6.7. An overview of the expected effects on cellular processes is shown in Figure 3.16.

A DNA fragment (-477 to +568) containing the HIV 5′ LTR was PCR amplified from the plasmid pLAI.BS (a gift from the Goyce lab) using PLAI FP (5′-TCTAGAACTAGTGGATCTTAG -3′) and PLAI RP (5′-GCTACAACCATCCCTTCAGAC-3′). An *in vitro* transcription reaction was performed in a 40 µL reaction containing 18 µL of HeLa nuclear extract in 20 mM HEPES, 80 mM KCL 3 mM MgCl 2 , 2 mM DTT 10 µM ZnCl 2 , 15 U rRNasin, 1 microg creatine kinase, 10 mM creatine phosphate, 250 microM of GTP, ATP, and CTP, 50 microM UTP, and 10 microCi [α- 32 P]-UTP. Reactions contained the PCR amplicon template (Sequence: Section C.2.7), Tat Protein (Sequence: Section C.2.6), and C-Terminal His_6 tagged TBP 6.7 (Sequence: Section C.2.3) were included in the concentrations given in Table 3.4.

Table 3.4: Reagent concentrations for TAT/TAR Dependent Transcription Assay

Reaction	[Template]	[Tat Peptide]	[TBP 6.7]
1	10 nM	2 µM	—
2	10 nM	2 µM	2 µM
3	10 nM	2 µM	0.2 µM

The reactions were incubated for 5 hours at 37 °C, and quenched by addition of 200 µL HSCB buffer (25 mM Tris-HCL, pH = 7.5, 400 mM NaCl). Following reaction stop 60 µg of glycogen was added to each reaction as a carrier, as was 1 µL of a 60 base radio-labelled RNA which functioned as a loading control. Proteins were extracted using phenol/chloroform/isoamyl alcohol, and nucleic acids were ethanol precipitated. Ethanol precipitated nucleic acids were resuspended in

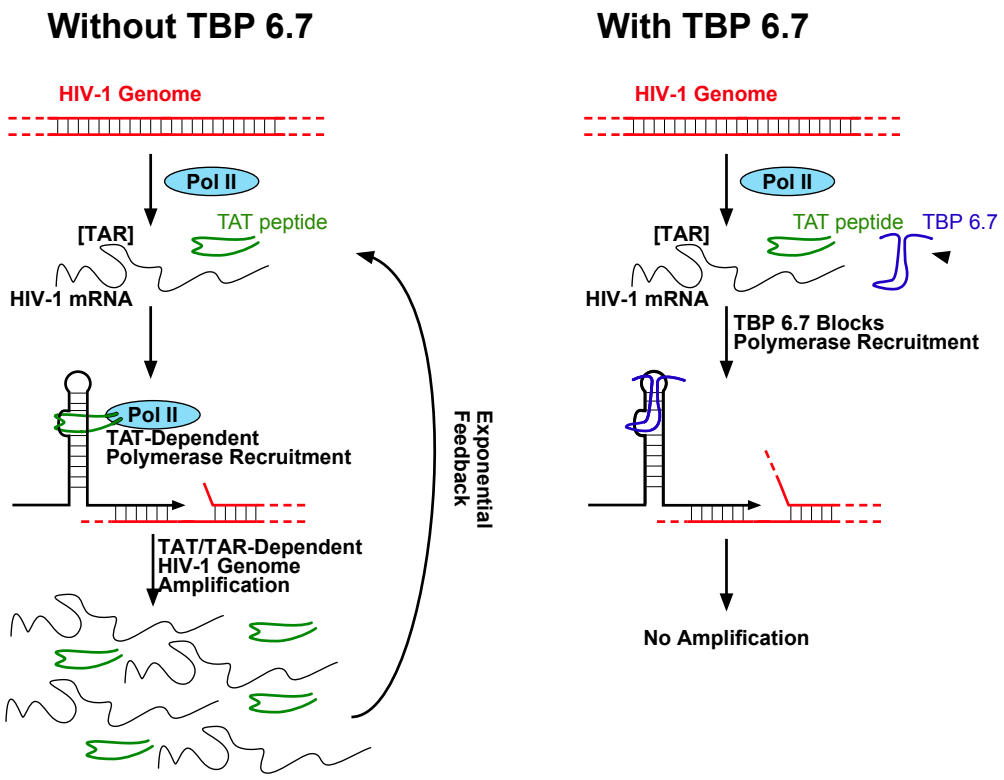


Figure 3.16: Biochemical Overview of Transription Assay Tat/TAR-dependent transcription results in production of mRNA copies of the HIV-1 genome. This transcription is blocked by the addition of TBP 6.7.

RNA loading dye, melted and refolded at 95 °C (to denature and prevent dimerization), and separated via PAGE. PAGE gels were developed using a phosphor imaging screen and a Typhoon imager.

A ~500 bp transcript was determined to be the key template-dependent transcript, and was quantified using ImageQuant software from GE Healthcare. The ~60 base spikant band was used to confirm that no significant variations occurred during the Phenol-Chloroform extraction and ethanol precipitation.

Gratifyingly, we observed a concentration-dependent suppression of Tat/TAR-dependent transcription in the presence of TBP 6.7. At the highest concentration of TBP 6.7 tested (2 μM), we observed virtually complete suppression of transcription (Figure 3.17).

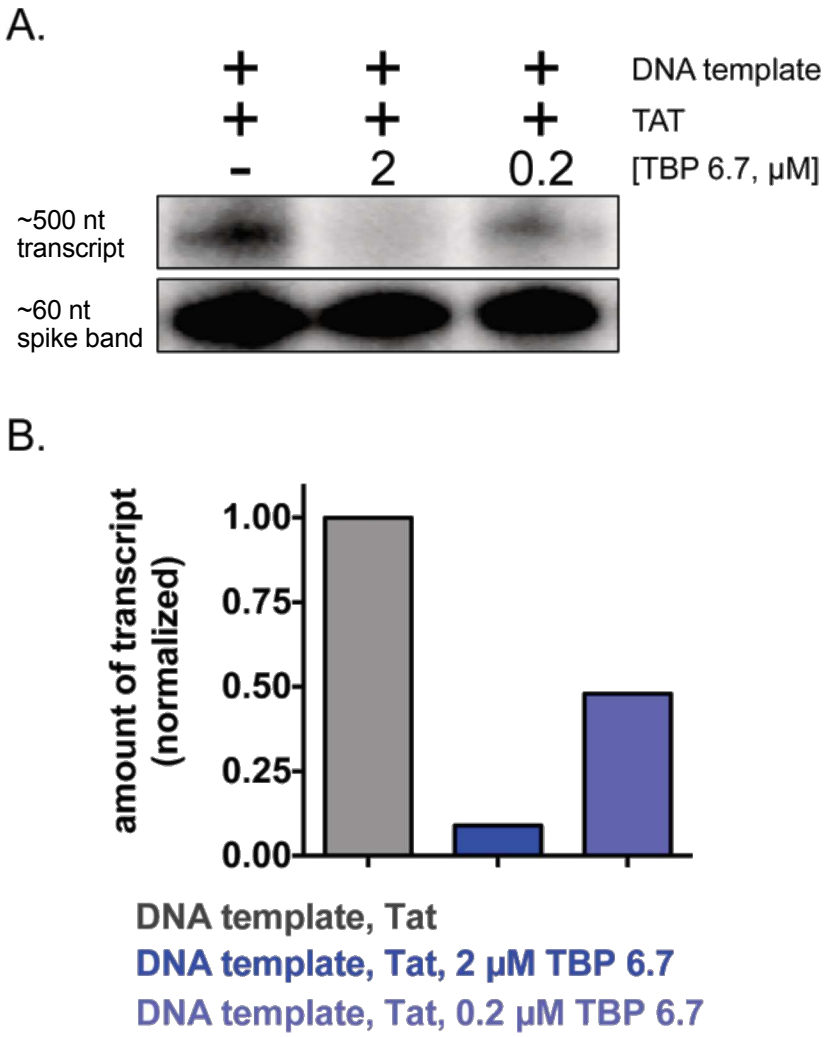


Figure 3.17: TBP 6.7 inhibition of TAT/TAR based Transcription TBP 6.7 was able to inhibit TAT/TAR based transcription in a concentration dependent manner. 2 μM of TBP 6.7 was able to all but turn off transcription, and 0.2 μM *reduces* transcription as can be seen in **A** a gel, and **B** the quantification of the transcription band. Adapted from [109].

3.9 Conclusions

The work described in this chapter marks the culmination of years of work in the McNaughton Lab, starting with Brett Blakeley's low-throughput screens of U1A point mutants. It also represents a major step forward for the field of RNA-binding proteins in general. The ability to engineer a protein which binds *an* RNA to bind a *class* of RNAs is useful, but engineering a protein which binds an RNA to bind a *different* RNA is even more so. There are very few such instances of a protein being engineered to have *altered*, rather than *broadened* specific for a target, especially for RNA binding proteins. TBP 6.7 especially fits this bill quite well, with its excellent (~500 pM) affinity for TAR (the tightest TAR-binding molecule, to my knowledge), while having negligible affinity for U1hpII.

The success of the protein in the functional transcription assay, as well as its ability to affect the entire bundled HIV 5′-UTR bode well for the possibility of using this protein (or peptides derived from it, see Chapter 5) to affect the HIV life cycle.

Our success in engineering and characterizing a binder with such excellent affinity was beyond our expectations. While we were excited and intrigued by the alterations to RNA binding based on the mutated TAR variants we used (Figure 3.9), we would ultimately require a crystal structure in order to *truly* understand the novel binding interaction we had developed.

Chapter 4

Crystallization of TBP 6.7 and TAR

4.1 Chapter 4 Introduction

4.1.1 Chapter 4 Summary

After yeast display affinity maturation (Chapter 2), and extensive characterization (Chapter 3), we determined that TBP 6.7 had excellent affinity for TAR and performed well in functional assays. Yet the chemical determinants of this interaction remained largely unknown. To learn about these determinants, we began a collaboration with Ivan Belashov in Joseph Wedekind's lab at the University of Rochester. I performed some basic confirmation that TBP 6.7 still bound TAR when it was changed back to the canonical U1A scaffold which would be used to find a crystal structure, and Ivan was able to obtain a crystal structure. This crystal structure illuminates the interaction between TBP 6.7 and TAR far beyond the crude mutagenesis data we gathered in Figure 3.9. Notably, we learned that rather than binding the single-stranded loop (as in the U1A–U1hpII interaction), TBP 6.7 binds TAR via the major groove of the double-stranded stem.

4.1.2 Chapter 4 Attribution

This chapter is adapted from [45].[3]

All crystallization research in this chapter performed by Ivan Belashov and Professor Joseph Wedekind of the University of Rochester Medical Center. My role was frequently active, but advisory, generally involving discussion regarding my typical methods of protein and RNA handling.

Molecular Dynamics simulations by Chapin E. Cavender and Professor David H. Mathews, also at the University of Rochester.

[3]Belashov, IA, Crawford, DW, Cavender, CE et al. Structure of HIV TAR in complex with a Lab-Evolved RRM provides insight into duplex RNA recognition and synthesis of a constrained peptide that impairs transcription. *Nucleic Acids Res*, 154:766–15, **2018**

I assisted with cloning of certain constructs, and performed initial checks to make sure that variants being used in crystallization studies maintained their ability to bind TAR (shown in Figure 4.1).

4.1.3 Chapter 4 Background

In Chapter 3, I and my collaborators had quantified the binding of TBP 6.7 to TAR. Additionally, we learned a great deal about the long-range effects of TBP 6.7 on the HIV-1 5′-UTR, and had determined that TBP 6.7 was able to disrupt the Tat–TAR interaction, and inhibit Tat/TAR-dependent transcription. Despite this knowledge of the *effects* of TBP 6.7 on TAR, our knowledge of the *mechanism* of interaction was limited to supposition based on the crude mutagenesis data shown in Figure 3.9. To obtain a crystal structure to better understand this interaction, we began a collaboration with the Wedekind Lab at the University of Rochester, hoping they could solve a crystal structure.

4.2 Preliminary Work

Though U1A was well-known to the Wedekind lab, there were some differences between the version they were accustomed to working with, and our own (our original wtU1A construct was a gift from the Laird-Offringa lab). The differences were relatively minor. Our lab's version has short linker sequences at the N- and C- termini and chemically minor changes at positions 75 (ours has a tyrosine, theirs a phenylalanine) and 88 (our version has an arginine, theirs a lysine), far from the putative binding face. These differences are highlighted in Figure 4.1A. No explanation was found for these small differences, and as they ultimately seemed to matter very little, no particular effort was made to find one.

To make sure that crystallization efforts were not doomed to failure from the start, I used ELISA to analyze the effects of removing the terminal linker sequences as well as the point mutations. As can be seen in Figure 4.1B, the truncations didn't affect binding appreciably, and the R88K only resulted in a minor difference. The Y75F mutation was more drastic, but given that

the binding is near saturation when the concentration is raised to a concentration of 100 nM, it was deemed to be suitable, and crystallization work was able to proceed.

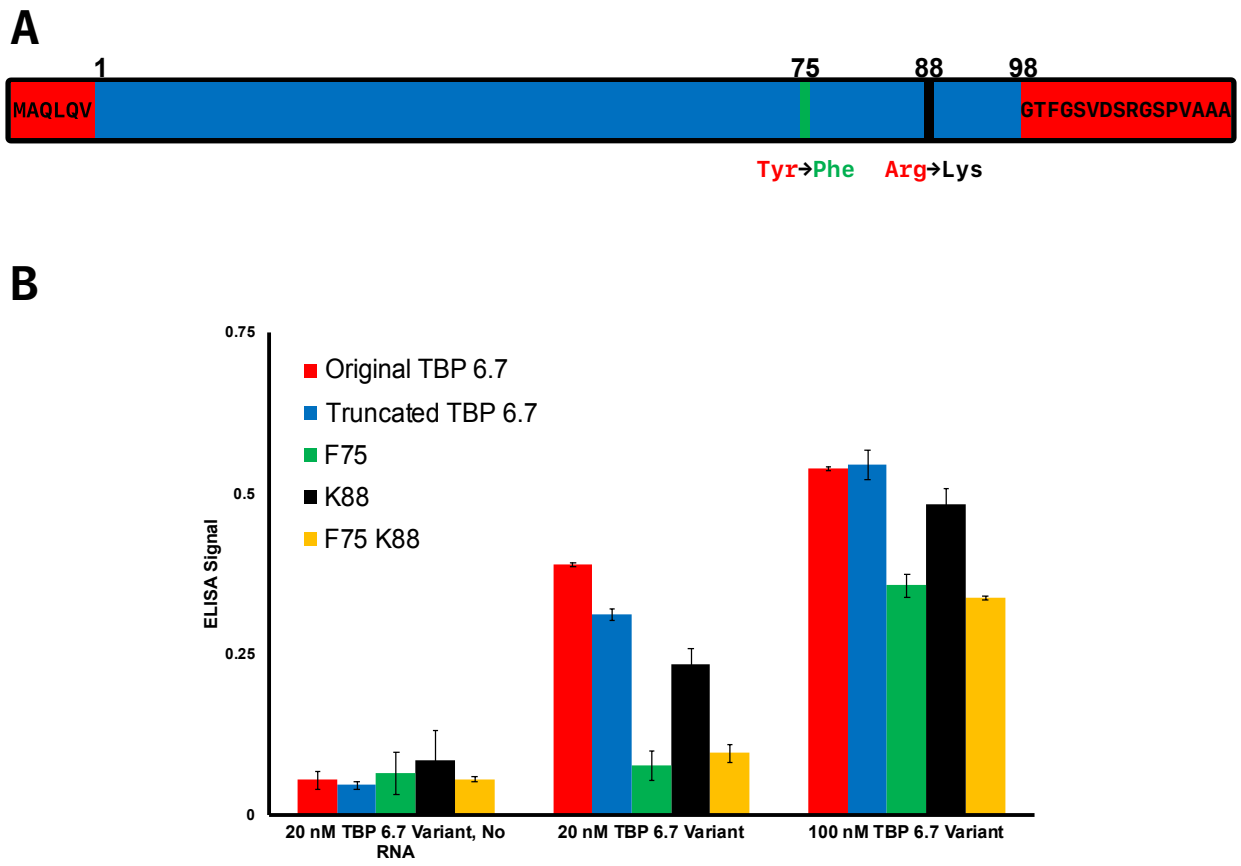


Figure 4.1: ELISA Assays to Analyze Variation Between U1A Scaffolds A The changes in U1A sequence between the construct in the McNaughton Lab and the Wedekind Lab needed to be analyzed.**B** The U1A version used by the Wedekind lab proved to be somewhat, but only somewhat, less avid when those mutations were added to TBP 6.7

4.3 Crystallization

Ivan Belashov was able to successfully crystallize, and solve the structure, of a TBP 6.7–TAR complex. This TBP 6.7–TAR co-crystal represents the first full crystal structure of the TAR element.

4.3.1 Protein Purification

TBP 6.7 was purified by Ivan Belashov in the Wedekind Lab in a similar manner to its purification in the McNaughton Lab.

TBP 6.7 was identified by the McNaughton Lab [109]. TBP 6.7 DNA was prepared as a synthetic gene (GeneScript Inc) comprising the human U1A sequence, yeast-display mutants [109], and Y31H/Q36R integrated for crystallization . After sub-cloning into pET28a(+) (Novagen) the crystallization [132]. After sub-cloning into pET28a(+) (Novagen) the thrombin site was modified by PCR to utilize TEV protease to cleave the N-terminal linker (ENLYFQ/G).

Point mutations were incorporated using the Q5 Site Directed Mutagenesis kit as described by the manufacturer (NEB) with primers from IDT. Protein expression in *E. coli* BL21(DE3) (NEB) was induced by 0.5 mM IPTG in LB at 20 °C. Cells were harvested after 4 h and pellets were frozen in N_2(l). Cells were thawed in a cell lysis buffer (CLB): 0.05 M Na-HEPES pH 7.5, 0.5 M NaCl, 0.02 M imidazole pH = 8.0, 0.0005 M EDTA, 0.005 M β-Mercaptoethanol and 0.01% (v/v) Brij35; the cell slurry was made 2 mg ml^{-1} in lysozyme (VWR).

After 20 min, cells were sonicated and the clarified supernatant was applied in batch to Ni-NTA resin (Pierce) equilibrated with CLB. After 2 h of nutation at 4 °C, resin was poured into a 1.5 cm × 10 cm gravity-flow column (CrystalCruz), washed with 40 column volumes of CLB, and two column volumes of wash buffer (WB): 0.05 M Na-HEPES pH 7.0, 0.3 M NaCl, 0.04 M imidazole pH 7.5, 0.005 M EDTA, 0.005 M β-ME and 0.01% (v/v) Brij35. Elution was in 3 ml fractions using elution buffer (EB): 0.15 M NaCl and 0.2 M imidazole pH 7.5. Fractions with 280 nm absorption were pooled and diluted with EB to a final imidazole concentration <0.02 M. TEV [133] was added (1:100 TEV:TBP) and the mixture was incubated at 4 °C. After 16 h, the reaction was incubated in batch with pre-equilibrated Ni-NTA, and supernatant was collected. Protein was loaded with an ÄKTA Pure (GE Lifesciences) at 0.5 ml min^{-1} onto a 5 ml HiTrap SP FF column (GE), followed by a linear gradient comprising: 0.15–0.85 M NaCl, 0.05 M Na-HEPES pH 7.0, 0.0025 M EDTA and 0.00025 M β-ME; TBP 6.7 elutes at ~70% as a sharp peak. The concentrated protein is polished on a HiPrep (16/60) Sephacryl S-300 HR column (GE Lifesciences). TBP 6.7 (M_r of 11.5 kDa) ex-

hibits higher retention than predicted by its M_r, eluting at or >1 CV. The yield is 2–3 mg/L of cells. Mutants were purified similarly.

4.3.2 Crystallization and X-Ray Data Collection

TAR RNA, produced by chemical synthesis, and purified by denaturing gel electrophoresis (Dharmacon), was suspended in 0.01 M Na-HEPES pH 7.5 to a concentration of 0.4 mM and heated at 65 °C. After 3 min, the RNA was diluted 10-fold with folding buffer (0.01 M Na-HEPES pH 7.5, 0.05 M NaCl and 0.002 M MgCl2) and incubated at 65 °C for 2 min. The RNA was cooled overnight to room temperature. TBP 6.7 was titrated drop-wise into folded RNA at a 1.2:1 molar ratio (48 μM protein to equal volume of 40 μM RNA) with vortexing.

The mixture was incubated at room temperature for 0.5 h and concentrated to 10–12 mg ml^{-1} based on 280 nm absorption using a Nanosep 3K Omega spin-filter (PALL); the final complex was 0.2 μm filtered (Millex, EMD). Crystals were prepared by vapor diffusion in which an equal volume of well solution (0.05 M Na-cacodylate pH 7.0, 0.1 M NaCl, 0.002 M $(NH_4)_2SO_4$ and 17% (w/v) of PEG-MME 5K) was added to 1.5 μl of TAR–TBP 6.7 complex with equilibration over 1 ml of well solution at 20 °C. Crystals grew within 72 h producing a half-octagon habit that reached 0.12 mm × 0.07 mm × 0.04 mm in 1 week. Cryo-protection was by serial transfer into well solution supplemented with 5–20% (v/v) glycerol followed by snap cooling in N_2(l). X-ray data were recorded at the Stanford Synchrotron Radiation Lightsource Table 4.1.

4.3.3 Phase Determination, Refinement and Analysis

The structure was determined by molecular replacement in PHENIX [134, 135] starting from U1A RRM1 [102] devoid of RNA. The initial TBP 6.7 model was generated by Phenix.autobuild [134], although TAR required manual building in Coot [136] with intervening cycles of Phenix.refine [134]. This iterative approach converged on $R_{cryst}/R_{work}/R_{free}$ values of 19.1%/18.9%/22.1% to 1.80 Å resolution (Table 4.1). An unbiased electron density map envelops all TAR nucleotides and the TBP 6.7 core (Figure 4.2A) indicating the quality of the refined structure. Reduced-bias omit maps demonstrate atomistic features that define placement of R47, R49* and

Table 4.1: X-ray Diffraction and Refinement Statistics of TBP 6.7–TAR Co-crystal

Data collection	
Space group	$P4_32_12$
Cell constants	
a = b, c (Å)	40.4, 284.6
α = β = γ (°)	90.0
Resolution (Å)	38.90–1.80 (1.83–1.80)
$R_{p.i.m.}$ (%)	2.6 (45.1)
CC1/2 (%)	98.7 (69.2)
I/σ(I)	19.9 (1.8)
Complete (%)	99.4 (91.8)
Redundancy	8.8 (7.9)
Refinement	
Resolution (Å)	37.2–1.80
No. reflections	23 297
R_{work}/R_{free} (%)	18.9/22.1
No. atoms	
Protein	746
RNA	572
Solvent	153
B-factors (Å^2)	
Protein	39
RNA	44
Waters	47
R.M.S. deviations	
Bonds (Å)	0.005
Angles (°)	0.759
Clash scored	0.4
Ramachandran (%)	
Allowed	100.0
Outliers	0.0
Coord. error (Å)	0.21

R52 side-chain rotamers and opposing bases (Figure 4.2B-D). These features are representative of the high-quality model that defines the TBP 6.7–TAR interface. The accompanying quality indicators (Table 4.1) provide confidence that the coordinates accurately describe the molecular details of protein-mediated TAR recognition. All cartoons, schematic diagrams and movies derived from coordinates were produced in PYMOL (Schrödinger, LLC). Cα superposition and Sc analysis were performed in CCP4 [137, 138].

4.3.4 Molecular Dynamics (MD) Simulations

MD simulations were conducted on the TBP 6.7–TAR complex, the TAR-(β2β3 loop) peptide comprising residues L41 to F59 (Figure 5.1), and isolated TAR RNA. The Amber 14 simulation package [142] was used to solvate crystallographic coordinates, or subsets thereof, in a box of OPC water [143] with 150 mM KCl. Starting coordinates were energy minimized using 500 steps each of steepest descent and conjugate gradient minimization with 25 kcal mol^{-1} $Å^{-2}$ positional restraints on solute atoms. Then, 10 cycles of alternating between minimization with decreasing positional restraints on the solute atoms and NVT dynamics were performed. After 250 ns of NPT equilibration, production dynamics simulations were performed using the AMBER ff14SB ([144–146]) force field in the NPT ensemble with periodic boundary conditions, a time step of 2 fs, and a direct space cutoff of 10.0 Å for nonbonded interactions. Bond lengths for covalent bonds involving hydrogen were constrained using the RATTLE algorithm [147]. Temperature was maintained at 300 K using a Langevin thermostat with a collision frequency of 1 ps^{-1}, and pressure was maintained at 1 atm using a Monte Carlo barostat. Simulations were performed on Nvidia Tesla K20X GPU cards. Six trajectories each of TAR–TBP 6.7, TAR–(β2β3 loop) peptide, and free TAR were each run for 4 μs for an aggregate time of 72 μs. For simulations of TAR–(β2β3 loop) peptide, the distance between the terminal carbon atoms of the β2β3 loop was restrained using a harmonic restraint with a force constant of 250 kcal mol^{-1} $Å^{-2}$ (roughly the strength of a covalent bond) to mimic peptide cyclization. Analysis of simulation interactions was performed using custom tools developed using the LOOS software [147].

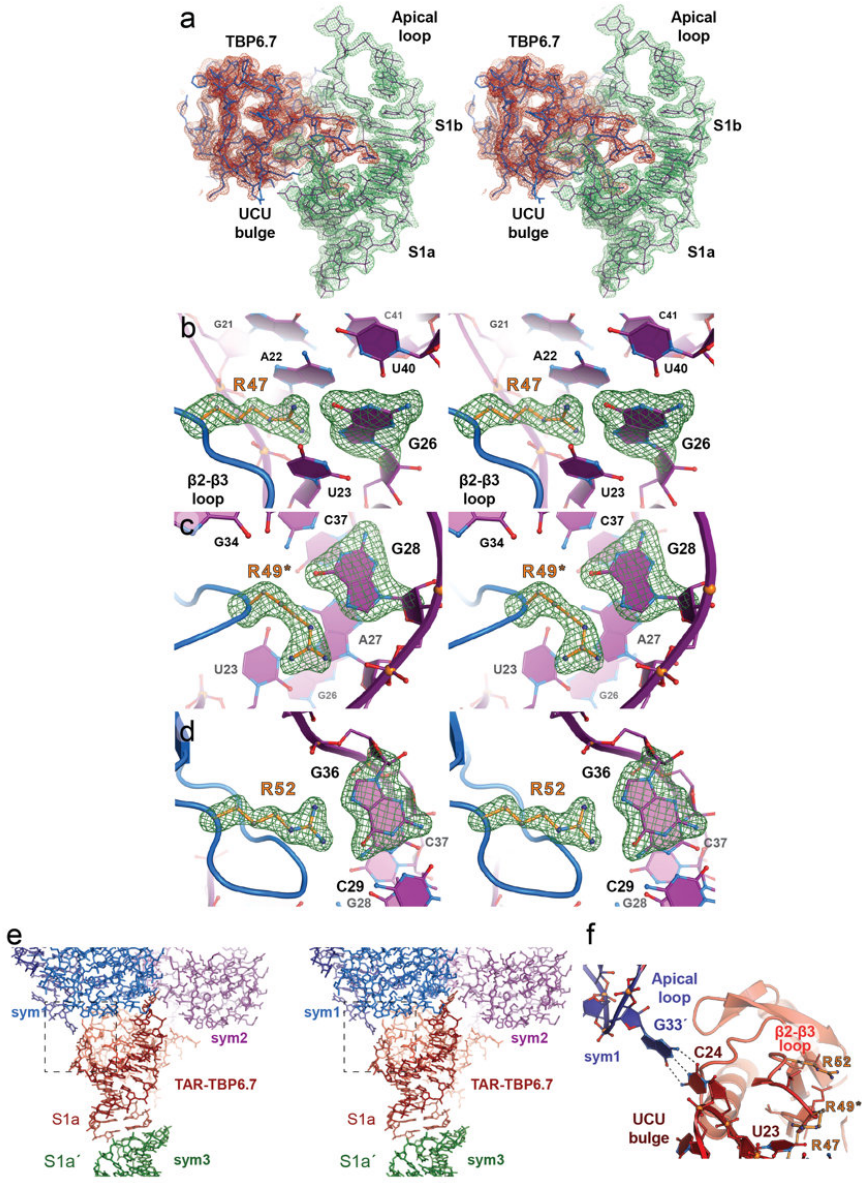


Figure 4.2: Electron Density Map of TBP 6.7–TAR Crystal StructureA An unbiased, composite iterative-build omit map [139] at 1.80 Å resolution contoured at 1.25σ. To differentiate the chains, the electron density surrounding the final refined coordinates of TBP 6.7 (blue bonds) is colored red, whereas that enveloping TAR (purple bonds) is green. The protein main-chain electron density is continuous from E5 to A95 with side-chain rotamers and carbonyl oxygens discernible for most amino acids. All nucleotides of the TAR 27-mer are well defined with obvious sugar puckers and base orientations about the N-glycosidic linkage. **B** Reduced bias simulated-annealing-omit (mFo–DFc) electron density map [140, 141] calculated from phases of the refined coordinates, but excluding atoms from the side-chain of R52 and base of Gua36; here and below, the map contour level is 3.0 σ and the final refined coordinates are depicted as ball-and-stick models. **C** Simulated-annealing-omit electron density map calculated from phases of the refined coordinates, but excluding atoms from the side-chain of R49* and the base of Gua28. **D** Simulated-annealing-omit electron density map calculated from phases of the final refined coordinates, but excluding atoms from the side-chain of R47 and the base of Gua26. **E,F,G** show the absence of symmetry contacts between TBP 6.7–TAR units. From [45].

4.4 Structural Analysis of the HIV TAR-TBP 6.7 Complex

4.4.1 Comparison to Previous Structures

To define the molecular details by which TBP 6.7 recognizes TAR, we determined the co-crystal structure (Table 4.1, Figure 4.2, and Section 4.3.2). In complex with TBP 6.7, TAR exhibits several architectural features consistent with solution studies of small ligands bound to the RNA. Hallmarks include stems S1a and S1b interrupted by the major-groove Uri23•Ade27-Uri38 triplex, flanked by a bulge that extrudes Cyt24 and Uri25 from its core (Figure 4.4A–D and Supplementary Movie S1). These characteristics are consistent with NMR analyses [55, 56] and persist on a μs timescale in our MD simulations (Figure 4.3A).

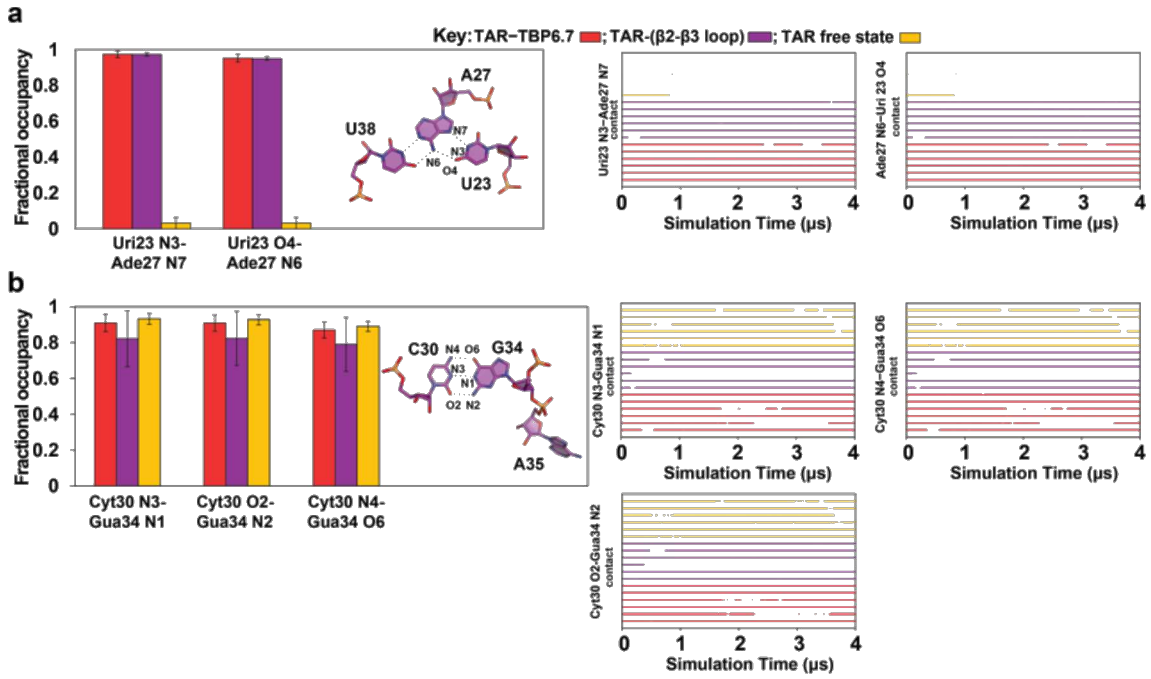


Figure 4.3: Fractional Occupancy and Molecular Dynamic Simulations for TAR-TBP 6.7 Complex Here and elsewhere, the fractional occupancy is reported for specific interactions indicated in the accompanying diagram (derived from the crystal structure). Hydrogen bonds are depicted as thin broken lines. Fractional occupancy (left) is the fraction of simulation frames—sampled at 1 ns intervals —in which a specific hydrogen bond is occupied. Bars represent mean fractional occupancy over the six MD trajectories ± SEM. Hydrogen bond occupancy is determined using a heavy-atom distance cutoff of 3.5 Å and an angle cutoff of 45° from linear. For each interaction, a time series of the occupancy during a given simulation (depicted as a colored line, right) shows a dot if the interaction is present. A color-coded key for each simulation type is shown at the top. In all cases, the molecules were pre-equilibrated for 250 ns. From [45], which also includes supplementary movie files.

Conversely, MD simulations conducted on apo-state TAR showed rapid dissolution of the triple (Table 4.3A, and Supplementary Movie S3 from [45]) concurring with ligand-free NMR analyses ([148–152]). Another hallmark of TAR is that the apical hexaloop interconverts between minor and major conformations [153]. In the latter, Uri31, Gua32 and Ade35 are flexible with adenine extruded [69, 152–154]. This is again mostly consistent with our co-crystal structure wherein Ade35 projects away from the hexaloop, whereas Gua32 and Uri31 stack on Cyt30 (Figure 4.4). Although unrepresented in solution ensembles of TAR-peptide complexes [55,56], the TAR/TBP 6.7 co-crystal structure exhibits a canonical cross-loop Cyt30-Gua34 pair (Figure 4.4C, E) supported by chemical modification experiments, NMR assignments, sequence conservation, and cyclin-T1 binding requirements [39, 68, 69, 155–157]. MD simulations indicate that Cyt30/Gua34 pairing is stable Figure 4.3, although transient dissolution and spontaneous reformation are seen for the TBP 6.7-bound and apo states. Nonetheless, the interaction appears to be a stable feature of the RNA conformational landscape (Supplementary Movie S2 from [45]). In one trajectory, Ade35 makes an excursion into the apical loop to displace Gua34 and interact with Cyt30 (Figure 4.3B, right, purple lines of trajectory four), agreeing with a low population state observed by NMR [153]. A likely site of conformational variation is extruded base Gua33, which forms a crystal contact with Cyt24 of the bulged loop from a neighboring molecule (Figure 4.2E, F). Neither base stacks appreciably inside the apical loop or bulged loop core on the timescale of MD simulations (Supplementary Movies S2 and S3 from [45]) and this contact does not influence TBP 6.7 binding. Figure 4.4

Comparison of the TBP 6.7 fold to that of U1A reveals that the evolved protein adopts the same mixed βαββαβ, architecture as parental RRM1 (Figure 4.4 A,F). A Cα superposition produced a modest rmsd of 1.1 Å, but local conformational differences are apparent. The greatest variations include the β2β3 loop (46–51, rmsd 3.9 Å) and the C-terminus (91–95, rmsd 3.6 Å), which were each subjected to saturation mutagenesis to achieve TAR binding [109]. When oriented similarly it is evident that TBP 6.7 and U1A engage their RNA targets in extraordinarily different ways (Figure 4.4A,F). Whereas TBP 6.7 binds TAR in the S1b duplex, U1A recognizes the distinctly single-

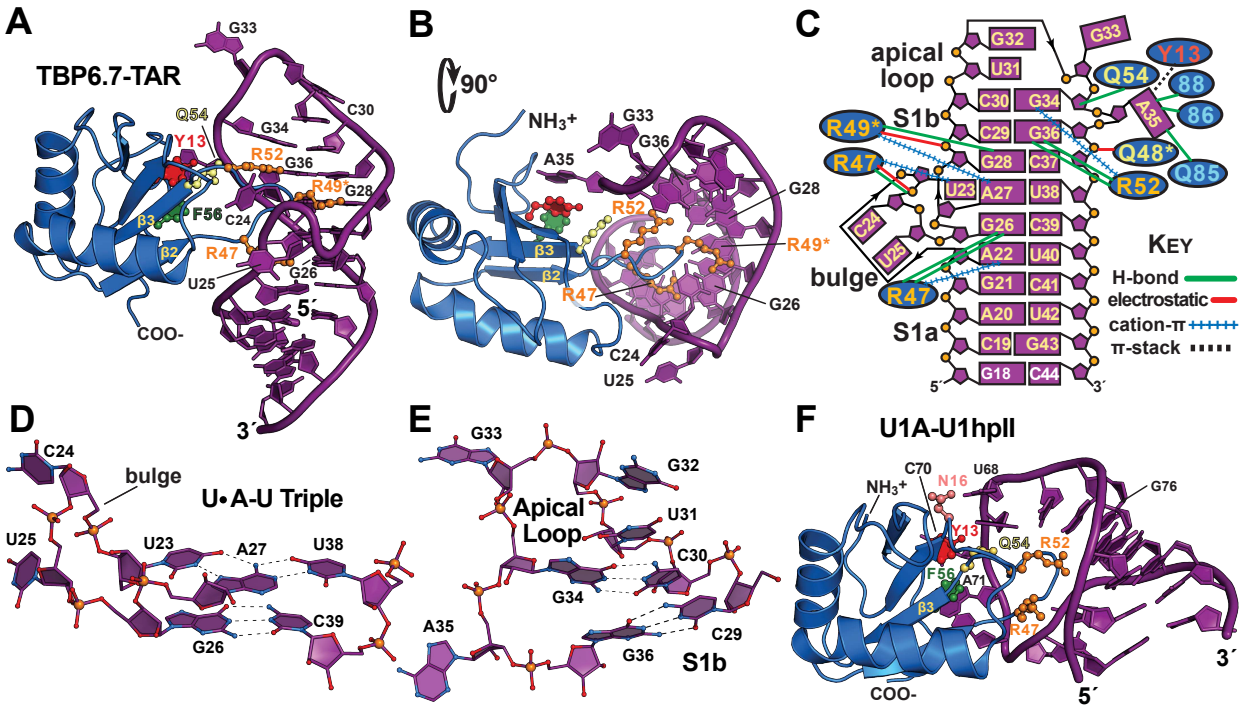


Figure 4.4: Overview of TAR Binding Protein 6.7–TAR complex Ribbon and schematic diagrams depicting the HIV-1 TAR–TBP 6.7 complex of this investigation and parental U1hpII-U1A. **A** Global view of the co-crystal structure depicting the TBP 6.7 RRM domain (blue) engaging TAR RNA (purple) in upper helical stem S1b. Arginines of the β2β3 loop that provide the principal determinants of TAR binding are depicted as ball-and-stick models (orange); similar depictions are provided for conserved RRM amino acids known as RNP2 (Y13) and RNP1 (R52, Q54 and F56). **B** Global view of the structure in A rotated +90°, providing a view looking through the apical loop and down the helical axis. The TBP 6.7 β2β3 loop penetrates deeply into the TAR major groove. **C** Schematic diagram depicting interactions between TBP 6.7 and TAR based on the co-crystal structure. Henceforth asterisks (*) indicate lab-evolved TBP 6.7 residues (see Figure 2.8). **D** Close-up of the TAR Uri23•Ade27-Uri38 major-groove base triple and the central bulge that interrupts stems S1a and S1b. Dashed lines joining ball-and-stick models represent putative hydrogen bonds unless noted otherwise. **E** Close-up view of the apical hexaloop and interface with the S1b closing base pair. **F** Global view of the U1hpII-U1A complex [102] oriented and colored as in **A**. U1A binds U1hpII primarily within the single-stranded region of the upper loop. From [45].

stranded loop of U1hpII between Ade66 and Cyt72 [102]. Despite fundamentally different modes of engagement, TBP 6.7 buries 1555 Å^2 in its protein–RNA interface, which is only 278 Å^2 less than the U1A–U1hpII complex. Importantly, the co-crystal structure reveals that numerous contacts to TAR originate in the β2β3 loop (Figure 4.4A–C), which yielded a clear consensus during selection that departs from U1A (Figure 2.1. Unexpectedly, the evolved C-terminus of TBP 6.7 is devoid of TAR contacts, implying that the minimal lab-evolved β2β3 loop is operative in the new mode of RNA binding, at least in this context.

4.4.2 TBP 6.7 Uses the RNP Motif to Recognize Double-Stranded RNA

Because TBP 6.7 maintains the classical RRM fold, we asked if it uses the conserved RNP motifs to bind double-stranded S1b of TAR, since these amino acids were unaltered in our approach [109]. This point is especially significant because RNP residues function classically in single-stranded RNA recognition [100]. In the U1A–U1hpII complex, RNA bases stack upon aromatic RNP side chains to provide affinity and recognition ([62,100,102,158,159]); Y13 of RNP2 and F56 of RNP1 stack on bases Cyt70 and Ade71 (Figure 4.4F and Supplementary Figure S3A). In contrast, Y13 of TBP 6.7 stacks on Ade35, but F56 does not engage TAR due to a lack of bulged bases flanking S1b (Figure 4.4A and Supplementary Figure S3B). Conversely, the Q54 amide of U1A RNP1 approaches the 2′-OH of Gua69 in U1hpII without interacting, whereas Q54 Nδ of TBP 6.7 hydrogen bonds to the 2′-OH of Gua34 in TAR (Figure 4.5A,B), consistent with its RNA readout role in other RRMs [100]. Finally, R52 of RNP1 recognizes the Hoogsteen edge of loop-closing pair Gua76-Cyt65 in U1hpII, as well as Gua36 in TAR (Figure 4.53C,D). The former interaction is the only instance of arginine-mediated base readout by U1A, although its simultaneous recognition of Ade66 N1 yields a non-optimal, inclined guanidinium-guanine interaction. A key finding is that TBP 6.7 still utilizes a subset of RNP amino acids to bind TAR, but affinity and specificity appear to arise primarily from the lab-evolved β2β3 loop, distinguishing it from U1A and other RRMs [100].

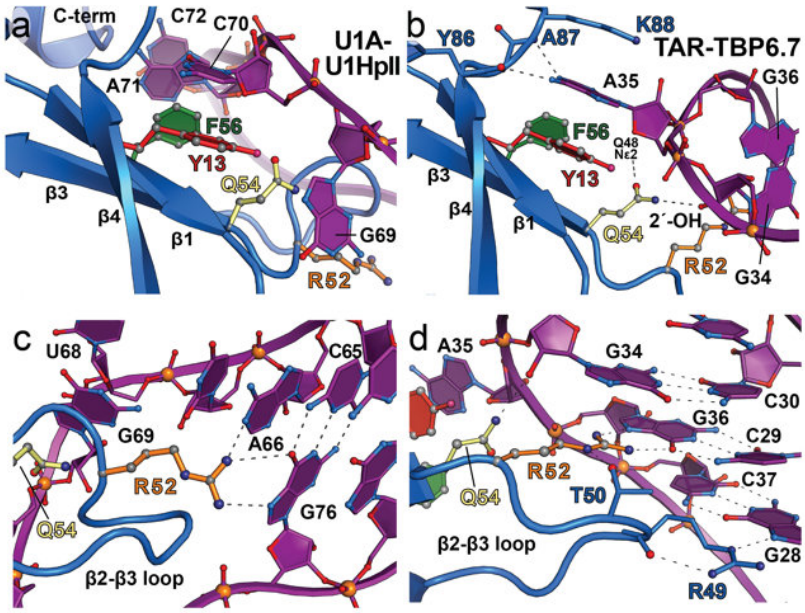


Figure 4.5: The RNP Motif in the TBP 6.7–TAR Complex The conserved RNP residues, a feature generally understood to be involved in *single*-stranded RNA binding participating in duplex interactions **A** View of the U1A RRM1 β-sheet face showing Y13 of RNP2 forming a π-stack with Cyt70 of the cognate U1hpII stemloop. F56 of RNP1 π stacks similarly with Ade71; Q54 of RNP1 makes no RNA contacts [102]. **B** View of TBP 6.7 (oriented as in **A**) wherein Y13 π stacks with Ade35; F56 makes no TAR contacts, but the Q54 amide hydrogen bonds to both the 2´-OH of Gua34 and the side-chain amide of Q48. **C** View of the U1A β2-β3 loop in which R52 of RNP1 uses its guanidinium group to read the N1 imino of Ade66 and the Hoogsteen edge of Gua76 located in the closing base pair of the U1hpII upper stem. The latter interaction is the only discernible site of duplex recognition by U1A. **D** View of the TBP 6.7 β2β3 loop oriented as in **C** to illustrate R52 recognition of the Gua36 Hoogsteen edge in stem S1b of TAR. From [45].

4.5 Thermodynamic Analysis

Thermodynamic analysis of the TAR/TBP 6.7 complex reveals that binding is enthalpy-driven (ΔH of –25 ± 0.2 kcal mol^{-1}) with an unfavorable entropy (–TΔS of 13.5 ± 0.2 kcal mol^{-1}) that yields a $K_{D,App}$ of 2.5 ± 0.1 nM (Table 4.2, Figure 4.7A). Analysis of the co-crystal structure suggested that binding interactions can be parsed into four groups: (i) arginines in the β2β3 loop that read guanine to impart specificity; (ii) β2β3 loop residues that interact with phosphate or 2′-OH groups; (iii) evolved protein–protein interactions that stabilize the β2β3 loop; and (iv) interactions outside the β2β3 loop. To test the energetic contributions of each, TBP 6.7 point mutants were prepared and evaluated by ITC for their ability to bind TAR (Table 4.2, Figure 4.7).

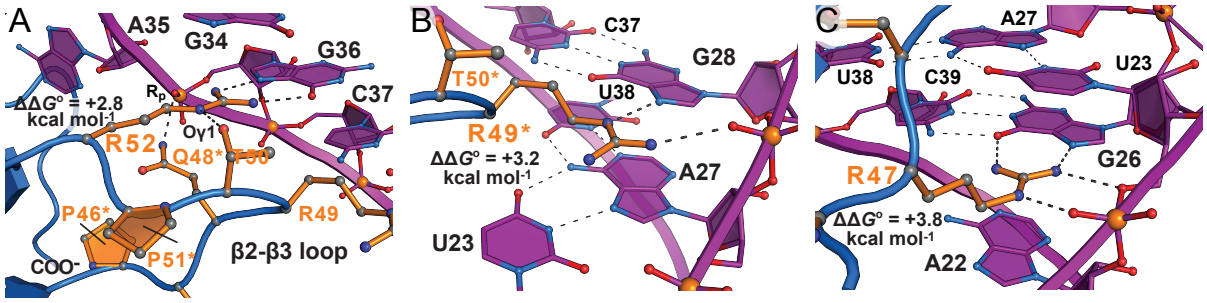


Figure 4.6: Detailed View of the β2β3 loop in the TBP 6.7–TAR Complex Close-up views of key interactions between the evolved β2β3 hairpin loop of TBP 6.7 and HIV-1 TAR based on the co-crystal structure. ΔG° values from ITC analysis of R-to-A mutations are taken from Supplementary Table S1. **A** R52 forms two hydrogen bonds to the Hoogsteen edge of Gua36; for clarity, some evolved amino acids in the β2β3 loop are omitted. **B** R49* forms a hydrogen bond with N7 of Gua28 and a salt-bridge to its non-bridging phosphate oxygen. **C** R47 forms two hydrogen bonds with the Hoogsteen edge of Gua26, as well as hydrogen bond and salt-bridge interactions to the Uri23 phosphate. Cation–π interactions and buried surface areas for each arginine are described in Figure 4.8. From [45].

4.5.1 Contributions of R52

Of the arginines in the β2β3 loop (Figure 2.1), R52 makes the fewest TAR contacts, making it straightforward to evaluate its binding contributions. Its guanidinium moiety donates hydrogen bonds from NH1 and NH2 to atoms N7 and O6 of Gua36 (Figure 4.6A and Supplementary Movie S4 from [45]), while forming a cation–π interaction with Gua34 of the apical loop (Figure 4.8A). Accordingly, the R52A mutation reduced binding by a factor of 116 (ΔΔG° of +2.8 kcal mol^{-1}) (Ta-

ble 4.2 and Figure 4.7B). R49* is the only arginine in the β2β3 loop that resulted from yeast display (Figure 2.2). This side-chain makes an equal number of contacts to TAR compared to R52, but the modes of interaction are different. The guanidinium group not only makes a hydrogen bond that recognizes N7 of Gua28, but also forms a salt-bridge to the nucleotide's *pro*-R_p oxygen while engaging in a cation–π contact to Ade27 (Figure 4.6, Figure 4.8 and Supplementary Movie S4 from [45]). Accordingly, R49A* yielded a larger ΔΔG° of +3.2 kcal mol^{-1}, corresponding to a loss in binding by a factor of *233* (Table 4.2 and Figure 4.7C).

Table 4.2: Thermodynamic Parameters for TAR-TBP 6.7 Binding at 20 °C

Sample	**KD,App**	**n**	**ΔH°**	**–TΔS°**	**ΔG°**	**ΔΔG°[1]**	**K_{rel}[2]**
TBP	nM	number sites	kcal mol^{-1}	kcal mol^{-1}	kcal mol^{-1}	kcal mol^{-1}	
TBP6.7	2.5 ± 0.1[3]	0.99 ± 0.02	–25.0 ± 0.2	13.5 ± 0.2	–11.6 ± 0.03	0	1
P46A	11.7 ± 2.5	0.97 ± 0.05	–22.7 ± 0.2	12.1 ± 0.1	–10.6 ± 0.1	1	4.7
R47A	1516 ± 163	0.96 ± 0.2	–7.5 ± 1.1	0.3 ± 1.1	–7.8 ± 0.1	3.8[4]	606.4
R47K	818 ± 61	0.91 ± 0.02	–10.5 ± 1.1	2.3 ± 1.1	–8.2 ± 0.03	3.4	327.2
Q48A	5.5 ± 1.0	1.00 ± 0.01	–22.6 ± 2.2	11.5 ± 2.1	–11.1 ± 0.1	0.5	2.2
Q48T	3.6 ± 1.9	1.00 ± 0.04	–25.7 ± 0.1	14.4 ± 0.2	–11.4 ± 0.3	0.2	1.4
R49A	583 ± 21	0.93 ± 0.1	–13.8 ± 0.7	5.5 ± 0.6	–8.4 ± 0.02	3.2	233.2
T50A	49.8 ± 19.1	1.00 ± 0.1	–21.6 ± 6.6	6.32 ± 1.9	–9.8 ± 0.2	1.8	19.9
P51A	10.8 ± 2.1	0.95 ± 0.05	–23.1 ± 0.1	12.4 ± 0.0	–10.7 ± 0.1	0.9	4.3
R52A	290 ± 57	1.00 ± 0.01	–14.3 ± 0.3	5.5 ± 0.2	–8.8 ± 0.08	2.8	116
Q54A	7.2 ± 2.4	1.05 ± 0.05	–21.9 ± 2.0	10.9 ± 1.8	–11.0 ± 0.2	0.6	2.9
ΔC-term	12.0 ± 3.0	1.01 ± 0.01	–24.7 ± 2.7	14.1 ± 2.5	–10.6 ± 0.2	1	4.8
TBP6.7-TAR2AP[5]	7.9 ± 0.2	1.00 ± 0.02	–25.9 ± 1.6	15.0 ± 1.6	–10.9 ± 0.01	0.7	3.2

[1] The difference of [ΔG° mutant – ΔG° TBP 6.7]
[2] Defined as the ratio of [mutant KD,App]/[wild-type KD,App] TBP6.7.
[3] Duplicate measurements were made for TBP6.7 wild-type and mutants with the exception of R49*A, which was measured in triplicate. Standard deviations of the mean are reported.
[4] Considered an estimate due to the low c-value associated with the measurement.
[5] Represents wild-type TBP6.7 injected into TAR RNA labeled at position 24 with 2-aminopurine (2AP).

4.5.2 Vital Contributions of R47A

Although R47 is present in the U1A sequence (Figure 2.1), it does not contact U1hpII RNA [102]. In contrast, R47 of TBP 6.7 makes the most extensive number of contacts with TAR forming an 'arginine fork' [160] wherein NH1 and NH2 hydrogen bond to O6 and N7 of Gua26, while Nϵ

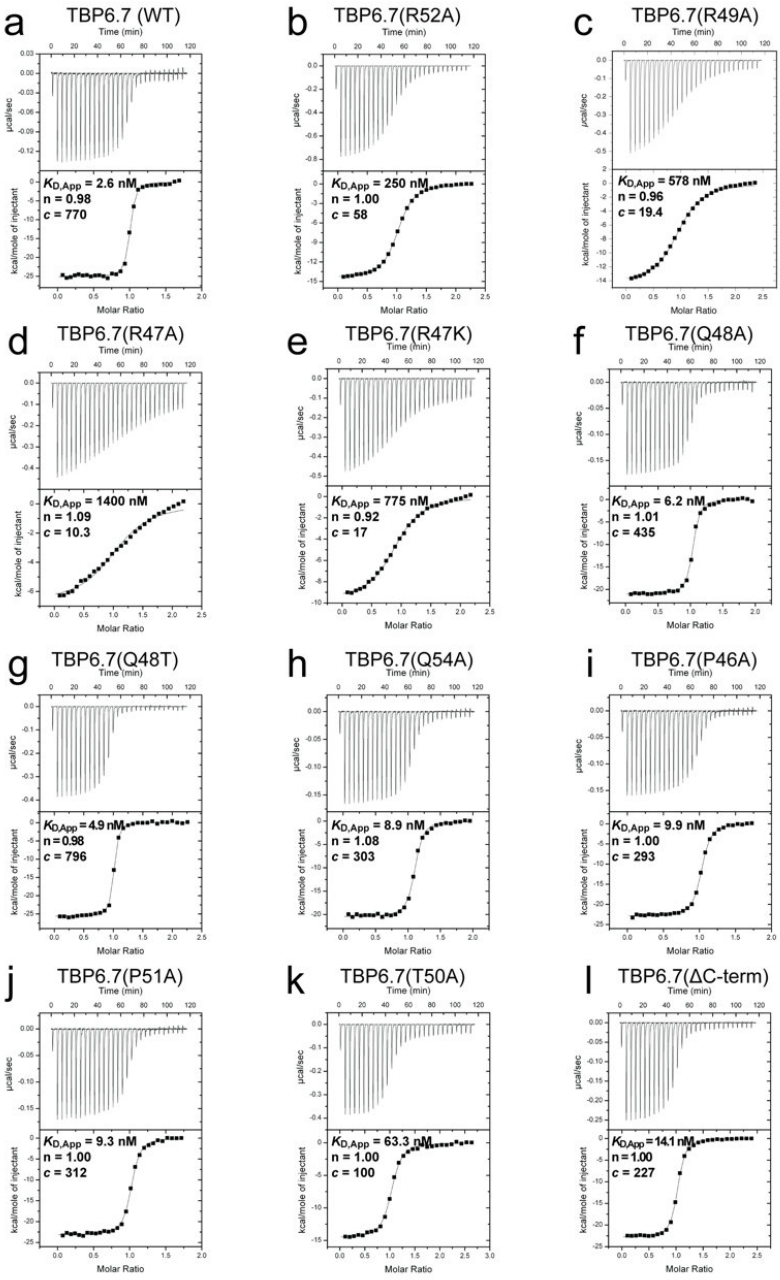


Figure 4.7: ITC Plots of TBP 6.7 Mutants Titrated into TAR The full isotherms and raw ITC data showing the interactions between varietions of TBP 6.7 and TAR. From [45].

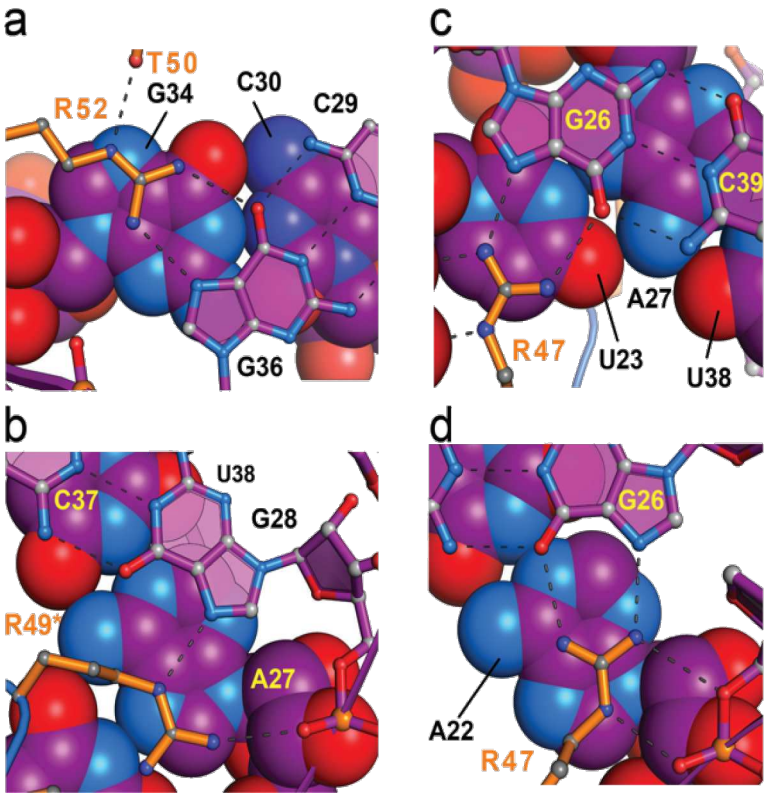


Figure 4.8: Schematic diagram of cation-π contacts between HIV-1 TAR bases and guanidinium groups contributed by the TBP 6.7 β2-β3 loop The figure cation-π interactions between the β2β3 loop and TAR. **A** R52 positions its Cζ atom near the π cloud of the Gua34, B) R49* positions its Cζ atom near the π cloud of the Ade27 imidazole ring, **C** R47 positions its Cζ atom near the π cloud of the Ade27 imidazole ring, and **D** R47 positions its Cζ atom near the π cloud of the Uri23 pyrimidine ring. From [45].

and NH2 hydrogen bond and salt-bridge to Uri23 O5′ and its pro-Rp oxygen (Figure 4.6C). The R47 guanidinium is sandwiched simultaneously between bases from Ade22 and Uri23 to form cation-π stacks (Figure 4.8C,D). As anticipated, R47A produced a large ΔΔG° of ~+3.8 kcal mol^{-1} corresponding to a loss in binding by a factor >600 (Table 4.2 and Figure 4.7D). The magnitude of this loss makes it tenuous to relate specific energetic contributions to the structure. An estimated 324 Å2 of buried area is ablated by this mutation—nearly double that of R52A (Figure 4.8). For a more conservative change, we examined R47K, which gave a ΔΔG° of +3.4 kcal mol^{-1} corresponding to factor of 327 in lost binding (Table 4.2 and Figure 4.7E). K47 could theoretically preserve salt bridge formation between its Nϵ and the Uri23 phosphate, as well as cation-π stacking, but hydrogen bonding to Gua26 and O5′ of U23 seem unlikely. From this analysis it is clear that R47 is of paramount importance for TAR binding, and that the positive charge of lysine is insufficient to attain optimal readout.

Our collective mutagenesis results support the crystallographic observations, revealing three tiers of TAR recognition corresponding to explicit modes of arginine readout with distinct free-energy profiles. MD simulations of the TAR–TBP 6.7 complex support the dynamics of the observed arginine–TAR interactions with higher maintenance of binding occupancy in more solvent-excluded regions Figure 4.9.

The simulations not only illustrate the feasibility of interactions to TAR in the context of full-length TBP 6.7, but also in the context of a minimal β2β3 loop peptide. An analysis of the other classes of interactions (ii) through (iv) (Section 4.5) demonstrated the roles of other evolved β2β3 loop residues in TAR recognition, their maintenance of a loop conformation productive for RNA binding, and the dispensability of the lab-evolved C-terminus for TAR readout. The co-crystal structure also provides a strong rationale for the binding affinities of various TAR mutants that were generated previously by the McNaughton Lab to probe sites of TBP 6.7 interaction with TAR RNA mutants [109].

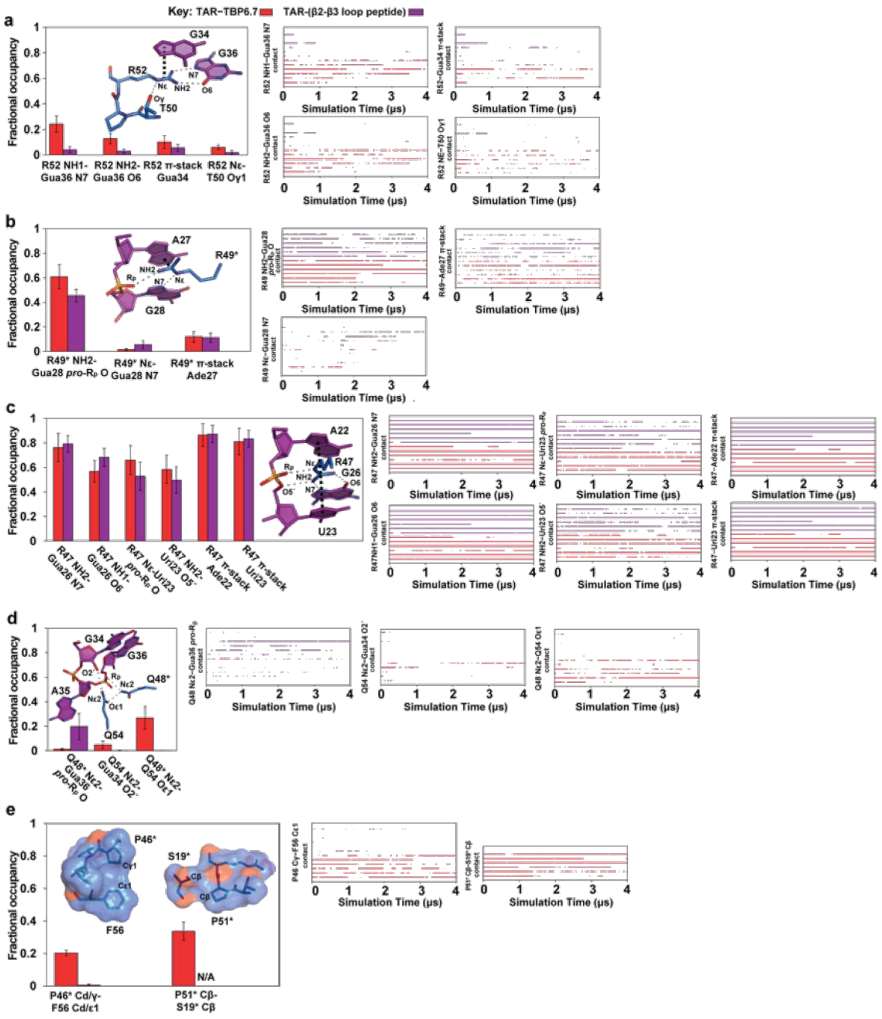


Figure 4.9: Fractional Occupancy Fractional occupancy (left) is the fraction of simulation frames, sampled at 1 ns intervals, in which a specific interaction is occupied. Bars represent mean fractional occupancy over six trajectories ± SEM. Hydrogen bond occupancy is determined using a heavy-atom distance cutoff of 3.5 Å and an angle cutoff of 45° from linear. Cation-π occupancy is determined using a distance cutoff of 4.5 Å between the centroids of the planar atoms of the nucleobase and those of the arginine guanidinium group, and an angle cutoff of 45° between the normal vectors of the planes defined by these atoms. For each interaction, a time series of the occupancy (right) shows a dot if the interaction is present. Associated graphed interactions are drawn as ball-and-stick models derived from corresponding regions of the co-crystal structure in which protein bonds are blue and RNA bonds are purple. Hydrogen bonds and salt bridges are depicted as thin broken lines; cation-π interactions are depicted as thick dashed lines. **A** The average occupancies of interactions to R52. These plots demonstrate that the interactions are dynamic, with a number of events showing loss and reformation of interactions across each of the simulations. **B** Average occupancy of interactions to R49*. **C** Average occupancy of interactions to R47. **D** Average occupancies of Q48* and Q54 interactions. E Average occupancies of van der Waals contacts between P46* to F56, and P51 to S19*. The van der Waals contact was set to a maximum C-C cutoff of 4.2 Å based on experimental distributions [161]. The van der Waals contacts are depicted here as semi-transparent surfaces for relevant groups of atoms; N/A means not applicable because the P51-S19* interaction is absent in the TAR-(β2-β3 loop) peptide complex. From [45].

4.6 Conclusions

4.6.1 General Conclusions

The results of the work described in this chapter have a number of important implications. The most obvious implication is the advanced understanding of the mechanisms of binding between TBP 6.7 and TAR, as well as the advanced *general* understanding of binding possibilities of binding between an RRM and a structured RNA.

The most notable implication of the advanced understanding of the *specific* mode of binding between TBP 6.7 and TAR is the prominence of the β2β3 loop in this binding interaction, making up ~⅔ of the buried surface area in the co-crystal. Most directly, this prominence inspires the peptide-based work in Chapter 5. Furthermore, there are less predictable, but obvious benefits of having an actual crystal structure of TAR (rather than the NMR based structures used to develop this work), not to mention a protein binding to it. Given that efforts toward a true *cure* for HIV infection would likely be based on some kind of transcriptional activation and suppression of HIV proviral transcription, the ability to affect and manipulate the TAR element is of utmost importance [162, 163].

In an abstract sense, this crystal structure represents a solution to this problem, that is to say: a particular set of molecular interactions that lead to binding of TAR. Knowledge of this set of interactions will surely inform the design or discovery of future TAR binders. This could mean that the direct contacts—notably the arginine trio discussed in Section 4.5—could be adapted. It could also mean less obvious motifs, such as P46 and P51 which make no direct contact, but P46 seems to guide residue F56 (on the RNP face) into place, and P51 affects S19 (itself a mutated residue).

Of more general interest is the fundamental mode of binding demonstrated by TBP 6.7, namely the double-stranded RNA binding based on major-groove recognition. The RRM class is generally, though not exclusively, made up of binders to *single-stranded* RNA. The fact of a new example of a dsRNA binding protein is itself noteworthy.

I hesitate to make concrete predictions, but it is inconceivable that this new, detailed data set won't be useful in the broader goals of developing binders to the TAR element, and general engineering of the RRM fold.

4.6.2 Relationship to Prior Work

Of particular interest to me is the fact that the C-helix has no obvious importance given that there are no apparent contacts between this region and TAR (seen in Figure 4.4), and the deletion of this region results in only a modest (~4.8-fold) decrease in affinity (shown in the ΔC-term data point in Table 4.2). Though unexpected (after all, the C-helix demonstrated *better* binding via yeast display than the β1α1 loop library in Figure 2.7, and in the ELISA data collected in Figure 3.4 there were clear differences in binding based *solely* on differing sequences in the C-helix. My explanation given in Section 2.6.7, that we may have been exerting selection pressure toward a C-helix which stays out of the way (rather than making any positive contributions), remains sound. I also posit that differences seen in the C-helix are more meaningful, perhaps as a cordon or stabilizer, in the context of solid-phase analysis (yeast display, ELISA, SPR) than they are in solution phase (ITC).

Realistically, the crystal structure shown in Figure 4.4 *supersedes* the mutagenesis data shown in Figure 3.9, but to me, of course, there is interest in the way the crystal structure *explains* the mutagenesis data. For instance, hp1, in which the UCU bulge is deleted (positions 23–25 on TAR) abolishes binding because the three vital arginines (R47, R49, and R52) are spatially distributed to accommodate this bulge, and R47 makes close contact with U23 on TAR. Obviously, these interactions can no longer occur in the bulgeless hp1. Realistically, this also makes most obvious case for the lack of binding to BIV TAR.

Of interest for a different reason are the case of hp2 (C30G/U31A)—which abolishes binding—and hp4 (G34C/A35U)—which results in reduced but not abolished binding to TBP 6.7, but does not significantly reduce binding to TBP 6.6. These changes may be most meaningful not due to any disruption between TBP 6.7 and TAR (though that is possible due to possible dis-

ruption of the G34-R52 cation-π interaction), but due to the fact that they disrupt the cross-loop base-pair which forms between C30 and G34 during TAR/TBP 6.7 binding. Ultimately, the hp2 and hp4 mutations were made with the assumption that they would not drastically alter the dynamics of TAR itself, but disruption to binding from these changes is in fact due to changes in TAR dynamics since the contacts between protein and RNA in these regions is minimal.

The apical loop mutations in hp3 (G32C/G33C) have little effect on TAR binding by TBP 6.7, and the crystal structure confirms the lack of contact between these apical bases and TBP 6.7. The stem mutants, hp5 (C29U/G36A) and hp6 (G28A/C37U/C29U/G36A), which decrease but don't abolish binding to TBP 6.7, are explained by the fact that R49 on TBP 6.7 recognizes the phosphate backbone of G28, as well as N7 on the nucleobase, both of which are still present on the A28 of hp5 and hp6. The contact between TBP 6.7 and TAR is a cation-π interaction between R49 on TBP 6.7 and G27 on TAR. This general interaction can still take place between R49 and the A27 on hp6.

Finally, the ultimate validation of our work, and the fundamental assumption that our "semi-design" strategy would be effective, was given in Section 4.4.1 in the comparison of the crystal structures of TBP 6.7 and the U1A from which it is derived. The rationale for using the U1A RRM as a scaffold was that the RNA-binding function of this RRM derived from its basic fold, and that it would be possible to maintain this basic function while changing the object of binding by making small, targeted mutations. Ultimately, there is a fundamental difference in function between U1A and TBP 6.7: high affinity and specificity binding to U1hpII RNA via a single-stranded RNA binding mode vs. high affinity and specificity binding to TAR RNA via a double-stranded RNA binding mode. But despite this fundamental alteration, both proteins maintain the canonical βαββαβ RRM fold, and a Cα superposition of TBP 6.7 and U1A has a RMSD of only 1.1 Å. This validation of a successful *alteration* in binding properties with minimal impact on overall structure is a triumph for the method of semi-design.

Chapter 5

Peptide Derivatives of TBP 6.7

5.1 Chapter 5 Introduction

5.1.1 Chapter 5 Summary

The co-crystal structure of the TBP 6.7–TAR complex revealed that the majority of the interaction (~⅔ of the buried surface) between the TBP 6.7 and TAR occurred at the β2β3 loop. We were intrigued by the possibility that this relatively small region of the protein may be independently sufficient for TAR binding, as it would have permeability and *in vivo* stability advantages over the full-length protein. Constrained peptides consisting of either the full β–turn–β motif, or only the loop residues, were synthesized and analyzed. The peptide binding to TAR was measured via a Fluorescence assay, and found to bind TAR with a K_D of 1.8 ± 0.5 μM. Furthermore this peptide was found to inhibit TAT/TAR-dependent transcription of the HIV DNA genome. When fused to a SUMO domain, the β2β3 motif selectively binds TAR over $(CUG)_{10}$ RNA as measured by ELISA. When displayed on bacteria (as an eCPX fusion) or yeast (as an Aga2 fusion), the peptide binds TAR. These successful fusion experiments indicate the possibility of being able to perform high-throughput screens in order to discover peptides with even better affinity for TAR, or other disease-relevant RNAs.

5.1.2 Attribution

This chapter is adapted from [45].[4]

ELISA (Figure 5.7), yeast display (Figure 5.9), and transcription assays (Figures 5.5 and 5.6) performed by myself.

[4]Belashov, IA, Crawford, DW, Cavender, CE et al. Structure of HIV TAR in complex with a Lab-Evolved RRM provides insight into duplex RNA recognition and synthesis of a constrained peptide that impairs transcription. *Nucleic Acids Res*, 154:766–15, **2018**

Bacterial display (Figure 5.8) performed by myself, with some assistance from Patrick Beardslee.

Polarization assay, the ITC data summarized in Table 4.2 and Figure 5.3B, performed by Ivan Belashov at the University of Rochester.

Molecular Dynamics simulations by Chapin E. Cavender and Professor David H. Mathews, also at the University of Rochester.

Peptide synthesis by Peng Dai, of the Pentelute Lab at the Massachusetts Institute of Technology.

5.1.3 Background

As described in Chapter 2, the decision to base the primary U1A-based library on the β2β3 loop was due to the large degree of contact between this loop and the RNA in the U1A–U1hpII interaction. We suspected that the β2β3 loop in the TBP 6.7–TAR interaction also accounted for disproportionate amount of the contact between protein and RNA. The co-crystal structure (described in Chapter 4) confirmed that the β2β3 loop accounted for ~⅔ of the buried surface area. This raised a question: could peptides based on the β2β3 loop (constrained to mimic the conformation on the TBP 6.7 protein) bind TAR on their own?

5.2 Synthesis of Constrained Peptide

Peptides were synthesized by Peng Dai in Bradley Pentelute's lab at the Massachusetts Institute of Technology.

Note: In the following text the NH_2 terminus and $CONH_2$ terminus will be designated, respectively, with the traditional "N" and "C" used for natural proteins, with the understanding that these designations correspond to the slightly different termini of a synthetic peptide.

The peptides were constrained by flanking the sequence in cysteines and forming a permanent perfluoroaryl covalent linkage between the thiol groups. The sequences used were either the full β–turn–β motif including the β2β3 loop (N-CLDILVPRQRTPRGQAFVIFC-C) for *pep-*

tide 1 or simply the β2β3 loop itself for *peptide* 1s (N-CVPRQRTPRGQAC-C). Structures of *peptide* 1 (the constrained β–turn–β peptide), and *peptide* 1s (the constrained loop only) are shown in Figure 5.1 with Liquid Chromatography–Mass Spec (LC-MS) analysis of the final products.

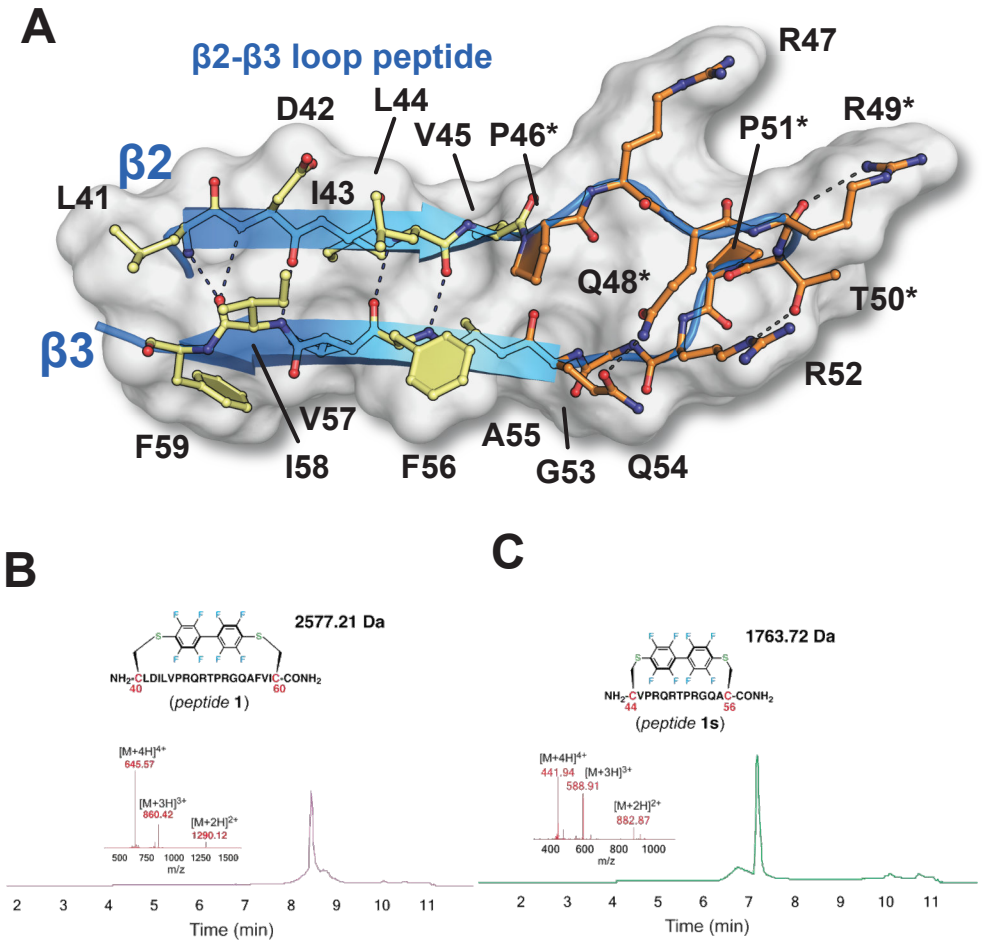


Figure 5.1: Structures and LC-MS Analysis of Constrained Peptides A Peptide 1 analysis showing a total ion current (TIC) chromatogram and the mass spectrum corresponding to the maxima point of the TIC peak (inset). **B** Peptide 1s analysis showing a total ion current (TIC) chromatogram and the mass spectrum corresponding to the maxima point of the TIC peak (inset). From [45].

5.2.1 General reagent information for synthesis of constrained *peptide* 1 and *peptide* 1s

1-[Bis(dimethylamino)-methylene]-1H-1,2,3-triazolo[4,5-b] pyridinium 3-oxide hexafluorophosphate (HATU) and Fmoc-L-amino acids were purchased from Chem-Impex International

(Wood Dale, IL). H-Rink Amide ChemMatrix resin was obtained from PCAS BioMatrix Inc. (Quebec, Canada). Peptide synthesis-grade N,N′-dimethylformamide (DMF), dichloromethane (CH_2Cl_2), diethyl ether and HPLC-grade acetonitrile were obtained from VWR International (Philadelphia, PA). Decafluorobiphenyl was from Oakwood Chemicals (West Columbia, SC).

5.2.2 Synthesis of Constrained *peptide* 1 and *peptide* 1s

Peptide **1** (N-CLDILVPRQRTPRGQAFVIC-C) and *peptide* **1s** (N-CVPRQRTPRGQAC-C) containing two free cysteines (Figure 5.1) were each synthesized on a 0.1 mmol scale on H-Rink Amide ChemMatrix resin. Solid-phase peptide synthesis was carried out on a synthesizer for automated flow peptide synthesis [164]. After completion, the resin was washed thoroughly with CH_2Cl_2 and dried under vacuum. The resin was transferred to a 50-ml plastic tube and the peptide was cleaved simultaneously from the resin while the side-chain was deprotected by treatment with 2.5% (v/v) water, 2.5% (v/v) 1,2-ethanedithiol and 1% (v/v) triisopropylsilane in neat trifluoroacetic acid (TFA) for 2 h at room temperature. The resulting peptide-containing solution was triturated and washed 2× with cold diethyl ether (pre-chilled at –80 °C). A gummy-like solid was dissolved in 50% H_2O:50% acetonitrile containing 0.1% TFA and lyophilized to yield the crude peptide. The peptide was reacted with decafluorobiphenyl in DMF for macrocyclization (43,44). The reaction mixture in DMF was quenched by water containing 0.5% TFA for 1:10 dilution, filtered and then purified by Reverse Phase HPLC (RP-HPLC). The solvent compositions for RP-HPLC purification were water with 0.1% TFA (solvent A) and acetonitrile with 0.1% TFA (solvent B). The diluted crude mixture was injected directly into an Agilent 1260 Infinity Automated LC/MS Purification System with a semi-preparative Agilent Zorbax 300SB C3 Reverse Phase-HPLC column (21.2 mm × 250 mm, 7 μm) operated with a linear gradient of 5–65% B over 82 min at a 4 ml min^{-1} flow rate. Fraction purity was assessed by LC–MS. Fractions containing pure, cyclized peptide were combined and lyophilized.

5.2.3 LC-MS Analysis of Constrained peptides

LC-MS chromatograms and associated mass spectra were acquired using an Agilent 6520 ESI-Q-TOF mass spectrometer. Mobile phases used for LC–MS analysis were: solvent C (0.1% formic acid in water) and solvent D (0.1% formic acid in acetonitrile). LC utilized a Zorbax 300SB C3 column (2.1 mm × 150 mm, 5 μm) with a column temperature set at 40°C and a flow rate of 0.8 ml min^{-1}. The gradient was: 0–2 min 5% D; 2–14 min 5–95% D; and 14–15 min 95% D. MS conditions were: positive electrospray ionization (ESI) extended dynamic mode in mass range 300–3000 m/z; temperature of drying gas equals 350°C; flow rate of drying gas equals 11 l min^{-1}; pressure of nebulizer gas equals 60 psi; the capillary, fragmentor, and octupole rf voltages were set at 4000, 175 and 750 V. LC–MS characterization of each peptide product is shown in Figure 5.1.

5.3 Preparation of TBP 6.7, SUMO, and SUMO-β2β3 Fusions for ELISA or Transcription

Protein and DNA sequences for TBP 6.7, SUMO, and SUMO fusions are provided in Section C.4. Briefly, plasmids containing indicated DNA sequences were constructed according the cloning procedure outlined in Section 3.2.2.

Cells were grown to confluence overnight in 5 mL cultures, and used to inoculate 0.5 L cultures of LB (Fisher) containing 100 μg ml^{-1} carbenicillin (GoldBio Technology) to an OD_{600} of ~0.6 and induced with 1 mM IPTG (Thermo Scientific) at 25 °C for 4–12 h. Cells were harvested by centrifugation (5000 × *g*, 10 min, 4 °C), resuspended in phosphate buffer (20 mM phosphate, pH 7.4, 0.15 M NaCl) prepared with cOmplete ULTRA Protease Inhibitor Tablets (Roche) and stored at –20 °C. For lysis, frozen cell suspensions were thawed and sonicated for 2 min. The lysate was cleared by centrifugation (9000 × *g*, 20 min, 4 °C) and the supernatant was mixed with 0.75 ml of Ni-NTA agarose (Fisher) for 10 min. The resin was sedimented by low-speed centrifugation for 5 min. Resin was washed with 30 ml of phosphate buffer containing 0.02 M imidazole, followed by a 10 ml wash with phosphate buffer containing 0.05 M imidazole. Proteins were eluted

using 2 ml of phosphate buffer containing 0.4 M imidazole. Eluted protein was dialyzed in 10K MWCO dialysis tubing (Thermo Scientific) against 2 L of phosphate buffer for ~12 hours, and then against a fresh 2 L of phosphate buffer for 4–6 hours. Purified proteins were quantified by absorbance at 280 nm using the calculated extinction coefficient. TBP 6.7 was prepared in an identical manner except that purification and initial dialysis were performed in HEPES buffer (10 mM HEPES, pH = 7.4, 50 mM KCl, 30 mM NaCl, 1 mM $MgCl^2$ and 1 mM EDTA).

5.4 Fluorescence Emission Analysis of TAR binding to *peptide* 1

5.4.1 Fluorescence Emission Assay

Due to the large quantities of material required by ITC, a fluorescence emission assay [165] was used to measure binding of *peptide* **1** to TAR.

Fluorescence measurements were conducted at 24 °C by titrating concentrated peptide in FL buffer (0.050 M HEPES, pH 7.5, 0.050 M NaCl, 0.050 M KCl and 0.002 M MgCl2) into 500 µL 100 nM TAR RNA 31-mer labeled with 2-aminopurine (2AP) at position 24 (5′-CGG CAG AU(2AP) UGA GCC UGG GAG CUC UCU GCC G-3′) known as (2AP)-TAR hereafter. The 2AP-RNA was purified by denaturing PAGE and folded as described (above). The excitation wavelength for 2AP was 320 nm and changes in emission were recorded at 390 nm as described [165] using a Fluoromax-3 fluorometer (Horiba Scientific). Data were fit to a one-site binding model, as described for TBP 6.7 binding to (2AP)-TAR (Figure 5.3).

5.4.2 Fluorescence Emission Results

The results of this assay indicate that *peptide* **1** binds TAR fairly well, with $K_D = 1.8 \pm 0.5$ µM, the binding curve can be seen in Figure 5.2.

One concern of this assay was whether the substitution of TAR for (2AP)-TAR was valid, especially since the substitution occurs in the critical bulge. This concern was alleviated by measuring the ability of TBP 6.7 to bind (2AP)-TAR. The results can be seen inFigure 5.3A, and indicate

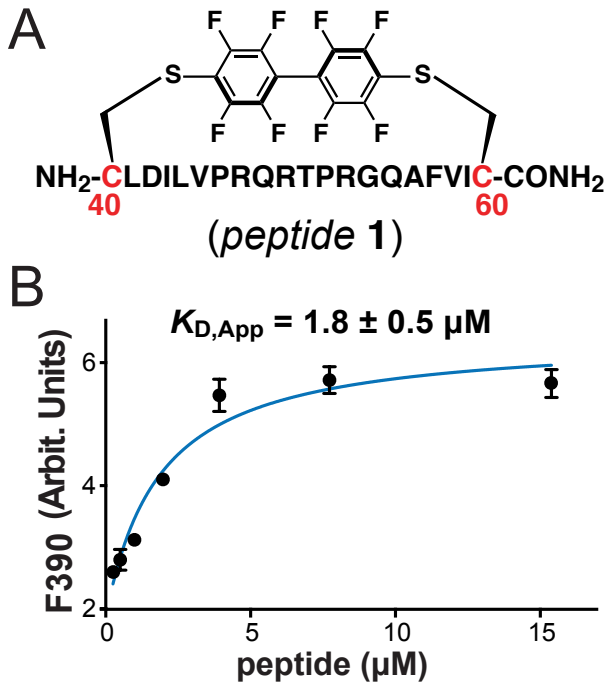


Figure 5.2: Fluorescence Assay Measuring Binding of *peptide* 1 to TAR The fluorescence assay measured binding to (2AP)-TAR of **a** *peptide* 1, with **b** showing the binding curve and the apparent K_D of 1.8 ± 0.5 μM. A similar assay showing binding of TBP 6.7 to (2AP)-TAR can be seen in Figure 5.3. From [45].

that though the use of (2AP) in place of the C at position 24 *may* have an effect, there is still fairly avid binding when measured by the fluorescence emission assay. with K_D = 10.8 ± 2.6 nM.

Additionally, these results were validated by an ITC experiment using TBP 6.7 and (2AP)-TAR, where it was found that, when measured by ITC, TBP 6.7 bound (2AP)-TAR with ~3-fold loss in affinity vs. normal TAR RNA (Figure 5.3B, Table 4.2), which indicates that the binding interaction between *peptide* 1 and TAR may in fact be tighter than measured.

Overall, these data indicate that our minimal β2β3 peptides appreciably bind TAR.

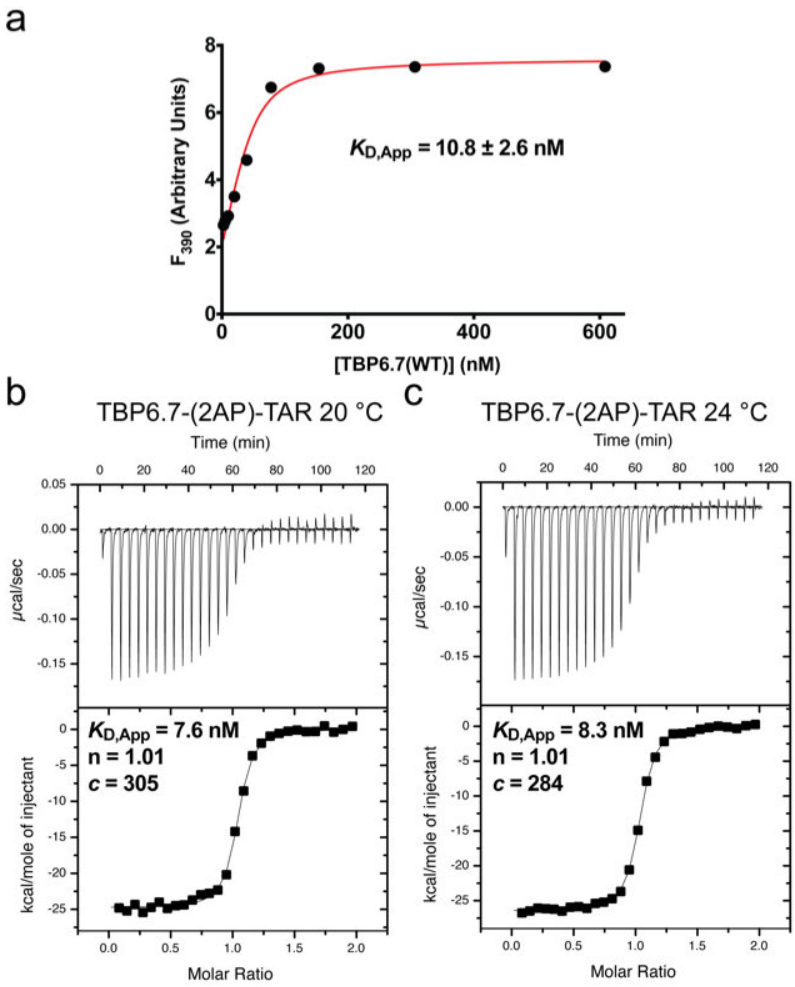


Figure 5.3: Fluorescence Assay Measuring Binding of TBP 6.7 to TAR Binding of (2AP)-TAR **a** via fluorescence assay with the binding curve shown. This was compared to **b** binding of TBP 6.7 to (2AP-TAR) as measured by ITC at 20 °C shown in **b** and **c** at 24 °C. Since this apparent K_D is ~10-fold worse than the binding of TBP 6.7 to TAR, it is possible that the 2AP affects binding, and *peptide* **1** may bind TAR with greater affinity than is shown by this assay. From [45].

5.5 ITC Inhibition Assays

I also tried to perform a similar experiment to that shown in Figure 3.15. The hope was that *peptide* **1s** would be able to disrupt the Tat–TAR binding interaction, and that that disruption in the interaction would be measurable via ITC.

5.5.1 ITC Inhibition Assay Methods

Assay Overview

These assays were performed in a similar manner to the assays described in Section 3.7.2. The same Tat peptide was used (N-RKKRRQRRRPPQGSQTHQVSLSKQPTSQPRGDPTGP KE-C), and the same TAR RNA (5′-GGC AGA UCU GAG CCU GGG AGC UCU CUG CC-3′). Experiments were performed at a variety of concentrations (3 μM or 10 μM TAR RNA), but the sample cell always contained either TAR RNA or TAR RNA pre-complexed with *peptide* **1s**, and the syringe always contained Tat peptide at 10x the concentration of the TAR RNA (20 μM, 30 μM, 100 μM).

Material Preparation

Since *peptide* **1s** was given to me as a lyophilized powder, I simply weighed out small aliquots of ~1 mg and was thus able to suspend the *peptide* **1s**, RNA, and Tat peptide in the same buffer (20 mM phosphate, pH = 7.4, 150 mM NaCl). Stocks were made at 1 mM of each. Though this does not, of course, take into account any leftover salts from peptide or nucleic acid synthesis, it did not seem to cause any obvious buffer mismatches. TAR RNA was also subjected to the melt-and-refold protocol described in Section 2.4.2, which is essential to its proper behavior in ITC.

Experimental Conditions

Tat peptide was either titrated into TAR RNA, or TAR RNA pre-incubated with a 2x or 10x molar excess of *peptide* **1s** or SUMO-β2β3 fusion. Pre-incubations occurred in volumes of ~400μL in 1.7 mL eppendorf tubes, rotating at 4 °C for 30–60 minutes.

ITC Conditions

ITC experiments were conducted using an ITC200 (MicroCal) using a 350 µL cell volume and a 75 µL syringe volume. All experiments consisted of 16 injections, an initial 0.4 µL injection and 15 injections of 4.98 µL. After a 60 second initial delay, injections occurred at 180 second intervals. Experiments were performed with a cell temperature of 25 °C, and a reference power of 2–5 µcal/sec.

Data were analyzed using Origin 7.0 (MicroCal, ITC200) using a "one set of sites binding model" for fitting. All data were reference subtracted by subtracting the mean heat of dilution from each data point.

ITC Results

Results of these ITC experiments can be seen in Figure 5.4, with full thermodynamic parameters given in 5.1.

Table 5.1: Thermodynamic Parameters for ITC, at 25 °C, of Tat Peptide Titrated into TAR RNA, with and without pre-complexing of β2β3 SUMO and *peptide* **1s**

Sample	$K_{D,App}$	**n**	**ΔH°**	**ΔS°**	K_{rel}
	nM	number of sites	kcal mol^{-1}	cal mol^{-1} deg^{-1}	
100 µM Tat into 10 µM TAR					
No Peptide	75.2 ± 14.7	1.47 ± 0.11	-7.7 ± 0.9	6.60	1
20 µM SUMO β2β3	101.8 ± 29.2	1.32 ± 0.02	-8.3 ± 0.2	4.22	0.74
100 µM *peptide* **1s**	255 ± 49.8	1.34 ± 0.02	-8.6 ± 0.2	1.20	0.29
30 µM Tat into 3 µM TAR					
No Peptide	16.1 ± 7.57	1.30 ± 0.0218	-9.0 ± 0.2	5.57	1
30 µM *peptide* **1s**	93.4 ± 41.4	0.63 ± 0.03	-9.4 ± 0.6	0.71	0.17

Though the data was not deemed sufficiently reliable to include in the final journal article due to the high error, and lack of clear saturation in most ITC curves (because of this, it is possible that the $K_{D,\,app}$ differences shown in Figure 5.4A are due simply to minor differences in heats of

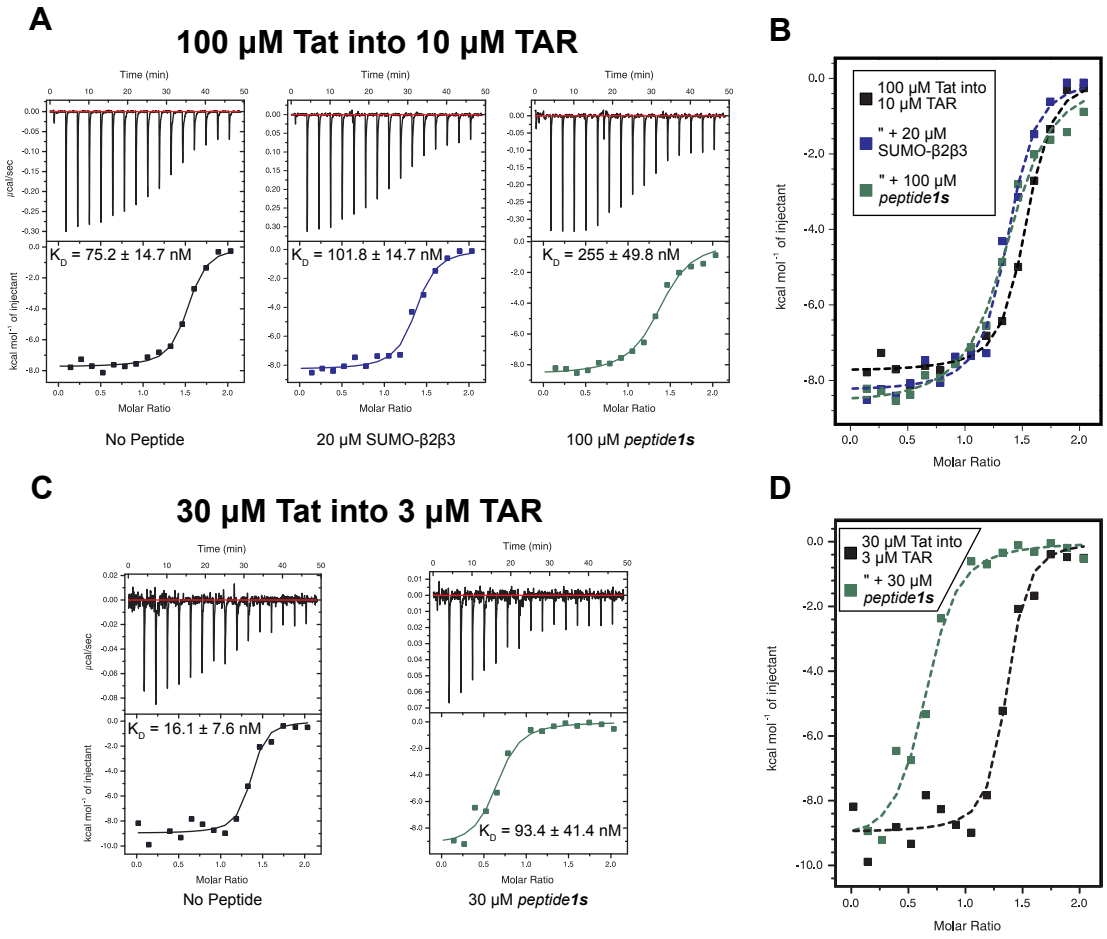


Figure 5.4: Inhibition of Tat–TAR Complex Formation by SUMO β2β3 loop and *peptide* 1s as measured by ITCShown in **A** are ITC data representing titration of 100 µM Tat peptide into 10 µM TAR RNA with either no peptide, 20 µM β2β3-SUMO, or 100 µM *peptide* **1s**. **B** shows the three binding curves overlaid. Shown in **C** are ITC data representing titration of 100 µM Tat peptide into 10 µM TAR RNA without and with 30 µM *peptide* **1s**. **D** shows both binding curves overlaid.

dilution at saturation). The data *do* indicate that sufficiently high concentrations of *peptide* **1s** can affect the ability of Tat peptide to bind TAR RNA. I feel it is worth including here, as it is somewhat corroborated by the more reliable (and more biochemically relevant) data shown in Section 5.6. This possible displacement via ITC is, incidentally most clear when 30 μMpeptide is used with 3 μMTAR RNA (Figure 5.4C,D), similar to the conditions used in the transcription.

5.6 Transcription Assay

5.6.1 Transcription Assay Methods

We then tested the ability of the shorter *peptide* **1s** (shown in Figure 5.1C) to target TAR using a known functional assay, the same transcription assay using HeLa Nuclear extract described in Section 3.8. The only major difference from the protocol described there is that rather than stopping the reactions with HSCB buffer and *subsequently* adding a loading control, the radiolabelled loading control (a purified RNA strand of either 180 or 350 nt) was mixed into the stop solution, which removed a potential source of error.

Each experiment consisted of 7 reactions (described in the caption of Figure 5.5. The count value of the ~500 nt band associated with Tat/TAR-dependent transcription was measured as a ratio of the count value of the control band, with the background subtracted based on densitometry values for areas adjacent to each band. This Tat/TAR band:control band ratio is considered the "absolute transcription." This value works well for comparing transcription levels *within* an experiment, but since there is so much variation in absolute magnitudes of transcription and spike band activity between experiments, it is not a valid comparison over multiple experiments, and performing three full experiments simultaneously is unrealistic. As such, the final value (the "transcript rel." y-axis value in Figure 5.6A) for each data point is given as the ratio of the absolute transcription value for each reaction and the uninhabited template + tat reaction. This relationship is illustrated in the following equations

$$Transcipt\ (absolute) = \frac{Densitometry500ntband}{Densitometry_{120ntband}}$$

$$Transcript\ (rel.) = \frac{Transcript(absolute)_{Reaction}}{Transcript(absolute)Reaction(Template+Tat)}$$

To give some idea of the background levels of transcription which occur in this experiment, and the variation between experiments, the uncropped gel from Figure 5.6B can be seen in Figure 5.5A, and another example gel is seen in Figure 5.5B. Visible on these gels is a ~180 nt band which, given its presence in all samples, is apparently a product of non-specific transcription. The fact that this band correlates very well to the control band is good confirmation that both *peptide* **1s** and TBP 6.7 operate by reducing Tat/TAR-dependent transcription, rather than being general transcriptional inhibitors.

5.6.2 Statistical Analysis

Unpaired, two-tailed t tests were performed with a Welch correction on data obtained from three separate transcription assays comparing untreated to inhibitor-treated conditions (Figure 5.6). The analysis was performed using Prism (GraphPad Software). The t values were: 4.82 (2) for 100 μM *peptide* **1**, 5.83 (2) for 10 μM *peptide* **1**, 4.07 (2) for 2 μM *peptide* **1**; and 8.55 (2) for 10 μM TBP 6.7. Parenthetical values indicate degrees of freedom.

5.6.3 Results

The statistical summary of three separate experiments, as well as a cropped-for-readability gel, is shown in Figure 5.6.

Efficient transcription from the HIV-1 5′-LTR requires an unfettered TAR–Tat interaction. Assays were conducted in HeLa nuclear extract to provide the endogenous transcription machinery, and exogenous Tat (Prospec Cat. No. hiv-129 417, Lot P1TATCB) was required for efficacious production of the 500 base transcript. Reactions lacking plasmid template and Tat, or without Tat, generated low levels of product (Figure 5.6, lanes I and II). In contrast, reactions containing template and exogenous Tat generated comparatively high levels of transcription product (Figure 5.6, lane III). When template, exogenous Tat, and various concentrations of *peptide* **1s** (100, 20 or 2 μM) were added, we observed concentration-dependent decreases in transcript production (Figure 5.6, lanes IV–VI). Statistically significant reduction occurred at 100 and 20 μM concentrations, and reduction at 2 μM was near statistically significant ($p = 0.085$).

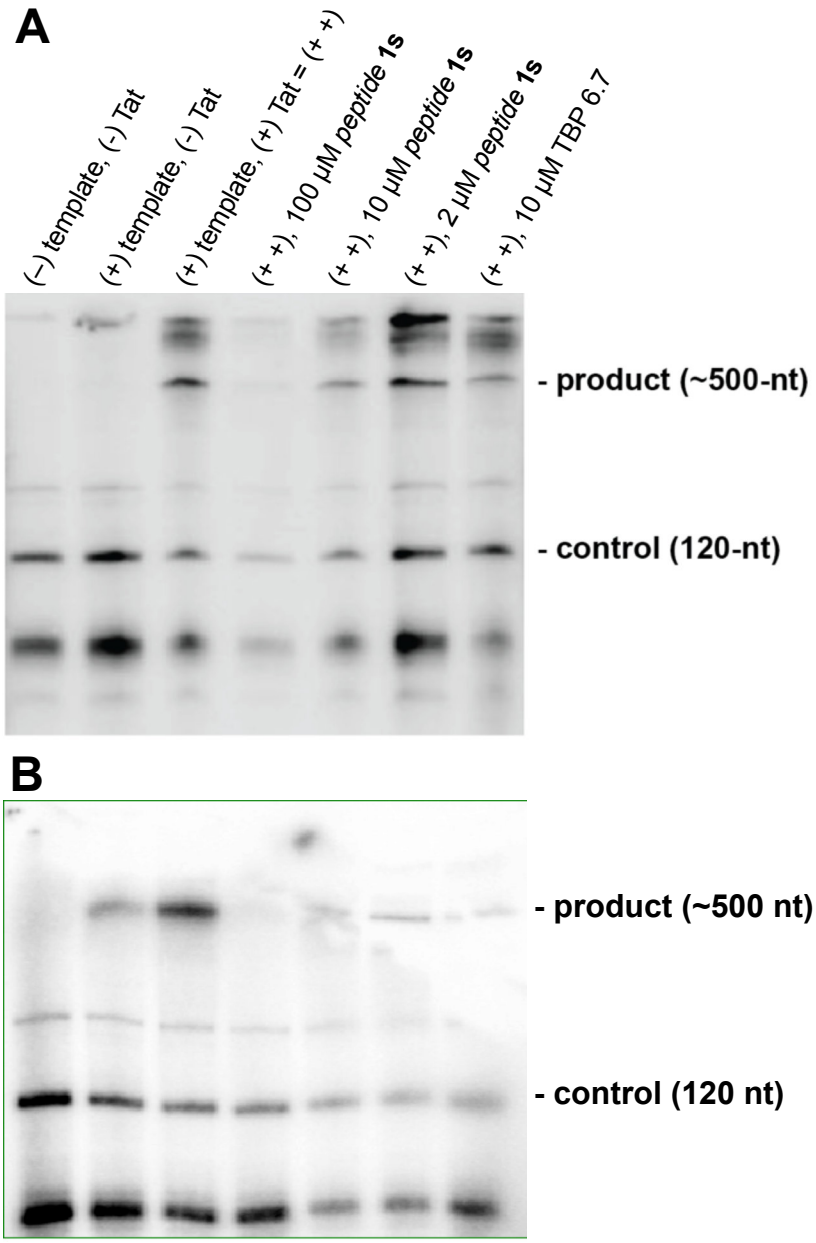


Figure 5.5: Transcription Assay Full Gels Two complete gels of independent transcription assays Figure 5.6 are shown in this figure.

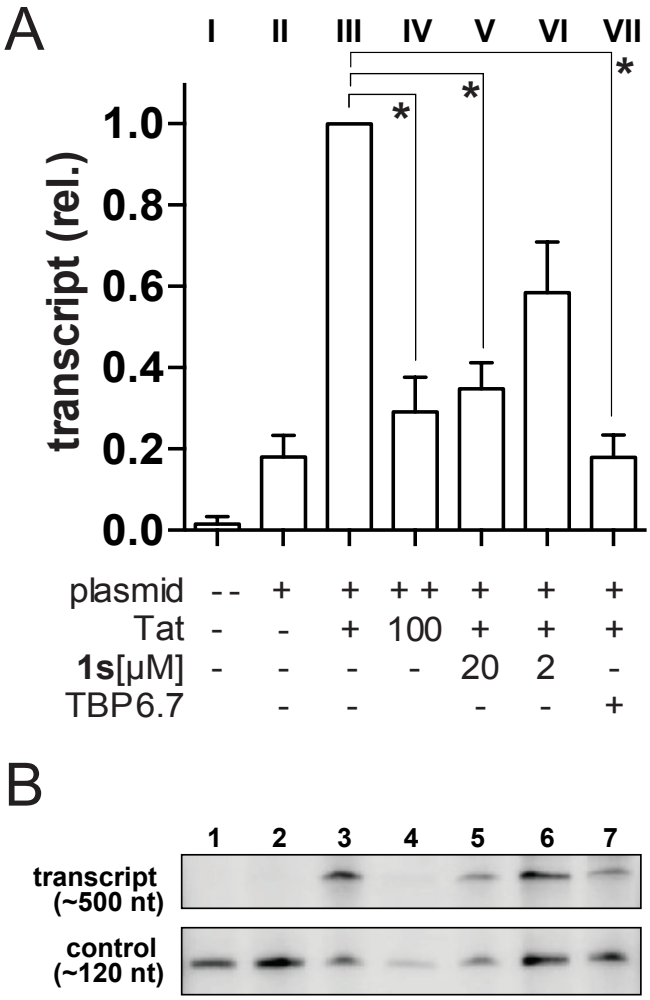


Figure 5.6: TAT/TAR Transcription Assay with *peptide* 1s In an *in vitro* transcription assay using HeLa nuclear extract *peptide* **1s** is able to inhibit TAT/TAR Dependent Transcription **A** with statistical significance at 100 μM and 10 μM, and near statistical significance at 2 μM. **B** shows a cropped-for-clarity version of the gel from Figure 5.5. From [45].

5.6.4 Transcription Assay Summary

The addition of 100, 20 or 2 μM of *peptide* **1s** (Figure 5.1B) resulted in approximately 70%, 65% or 40% suppression of transcription product (Figure 5.6, lanes 4–6 and Figure 5.5). Consistent with our previous findings [109], 10 μM TBP 6.7 inhibits TAR–Tat-dependent transcription (Figure 5.6, lane VII and Figure 5.6, lane 7). The results imply that *peptide* **1s** mimics the β2β3 loop of TBP 6.7 and serves as a minimal TAR recognition peptide capable of restricting an essential viral activity.

5.7 SUMO Fusions of the TBP 6.7 β2β3 Loop

Once it was established that peptides derived from the β2β3 loop of TBP 6.7 could serve as TAR binders and as functional inhibitors Tat/TAR-dependent transcription, I attempted to create a platform that could be used to synthesize these peptides recombinantly in *E. coli*. The efficacy of these fusions was analyzed via ELISA.

5.7.1 ELISA Protocol

ELISA was performed with a condensed version of the protocol described in Section 3.3.1 involving a co-incubation of protein and RNA.

The solid-state scaffold was a clear, 5 picomole well^{-1} streptavidin-coated 96-well plates (Pierce). The plate was pre-incubated for 1 h with wash buffer (20 mM phosphate pH 7.4, 150 mM NaCl, 0.05% Tween-20 and 0.1 mg ml^{-1} BSA). During pre-incubation 100 μl of TAR (5′-GGC AGA UCU GAG CCU GGG AGC UCU CUG CC-3′) or CUG10 (5′-CCG CUG CUG CUG CUG CUG CUG CUG CUG CUG GGC-3′) RNA modified with a 5′-biotin (IDT) was incubated in 100 μl of buffer with 1 μM of either SUMO-β2β3 variant or SUMO for 1 h, rotating at 4 °C. The pre-incubation buffer was removed from the ELISA plate and the RNA–protein mixture was incubated on the plate for 2 h. Wells were then washed 3× with 200 μl of wash buffer with shaking for 5 min. Next, a 1:10,000 dilution of HRP-conjugated anti-FLAG antibody (Abcam, ab2493) was made with Odyssey Blocking Buffer (Li-Cor) and 100 μl was incubated in each well for 30 min

at 25 °C; each well was then washed 4×. Colorimetry was developed for 20 min using 100 µl of TMB-One substrate (Promega). Absorbance was measured at 655 nm on a plate reader. ELISA experiments were repeated in triplicate.

5.7.2 ELISA Results

The results of the ELISA comparing the binding of SUMO-β2β3 to TAR and $(CUG)_{10}$ are shown in Figure 5.7A. As expected, a SUMO domain without attached peptide shows comparatively low levels of binding. In contrast, SUMO-β2β3 binds TAR, and does so with greater affinity than it binds $(CUG)_{10}$ (Figure 5.7). The results collectively demonstrate that the lab-evolved β2β3 loop retains TAR binding outside the context of TBP 6.7 when presented as a fusion protein.

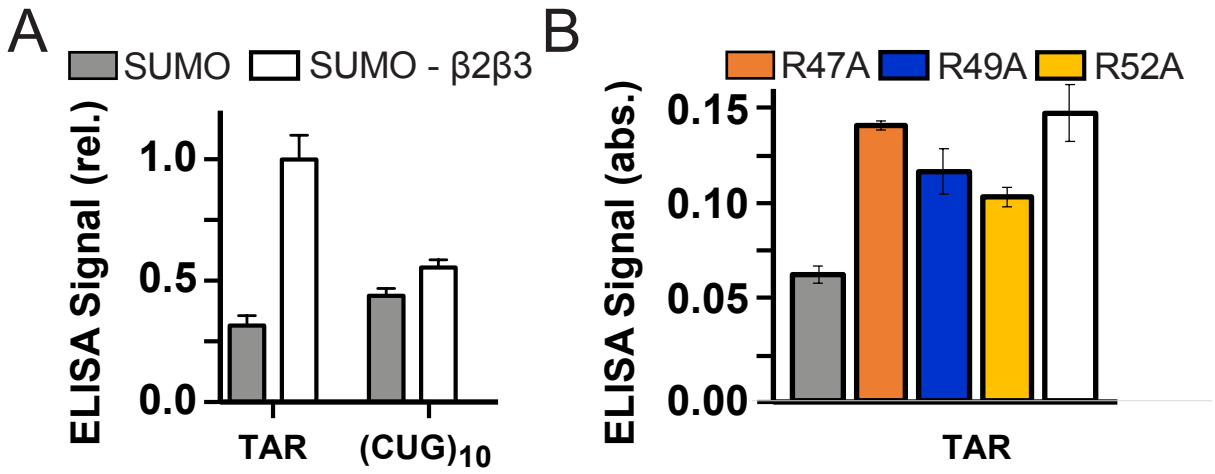


Figure 5.7: ELISA Data Showing Binding of SUMO Fusions of the TBP 6.7 β2β3 Loop, and associated Arg → Ala mutantA A SUMO-β2β3 loop fusion binds TAR, but not $(CUG)_{10}$ and **B** Shows the affects of loop mutations on this binding. Interestingly, the keystone R47 seems to be the most readily removable of any of the trio. Adapted from From [45].

The results comparing the evolved β2β3 to arginine mutations of the same are more puzzling (See Figure 5.7B). Removing the keystone R47 residue and replacing it with an alanine has only minimal effect on TAR binding. In fact, replacing *any* of the three important β2β3 Arg residues has only a minimal effect on binding. This is in contradiction to the ITC data, in which replacing any of these residues has deleterious effects on binding, with TBP 6.7 R47A essentially losing binding ability (Figure 4.7, Table 4.2).

My own supposition is that since the R47 residue was *not* randomized in the selection experiment, yet is absolutely *key* to binding TAR, the β2β3 loop evolved largely to *position* this arginine. As I discuss frequently, the challenge of engineering RNA binders is that there is only a limited diversity of function (charge/charge, cation/π, π/π, hydrogen bonding), and that specificity must come from *placement*. It is possible that only in the context of the full RRM is the R47A so perfectly placed, and that in the less sterically-hindered and more flexible context of the β2β3 peptides, the other Arg residues are able to form their own binding interaction network, albeit a less effective one. If this is the case, it is likely that a double or triple Ala for Arg substitution *would* abolish binding.

One important caveat is that these ELISA data were performed at the very edge of the capability of the assay, and that the absolute absorbances of ~0.15 are ~30-fold lower than the absolute absorbances of the ELISA assays discussed in Section 3.3.1, either due to significantly reduced absolute binding or (my own hypothesis) decreased availability of the FLAG tag for binding the HRP-conjugated anti-FLAG antibody. With that said, recent unpublished data based on eCPX bacterial display (similar to the data discussed in Section 5.8.1) indicate that peptide binding is not abolished by the loss of any given arginine.

5.8 Surface Display Assays

To test the ability of the peptide to express and function in a context in which it could later be subjected to screening, we fused the full β–turn–β motif to either eCPX (for display on *E. coli*, or Aga2 (for display on *S. cerevisae*).

The yeast display assay using Aga2 is discussed at length in Section 2.4.2. The bacterial display assay was performed using a system developed in the Daugherty lab [166, 167]. The system utilizes a fusion to a modification of outer membrane protein X (ompX), an *E. coli* membrane protein amenable to fusion. The version used for fusion proteins here is a modified version, enhanced circularly permuted membrane protein X (eCPX), which is amenable to fusions, resulting in display of ~10^5 copies of the fused peptide or protein on the surface of an *E. coli* cell.

The version of eCPX used here is the result of refinement by Angeline Ta in the McNaughton Lab, and includes a multiple-cloning site which enables easy fusion with both the eCPX protein and a *myc* tag

5.8.1 Bacterial Display

The full β2β3 loop peptide sequence of TBP 6.7, analogous to *peptide* **1** (N-LDILVPRQRTPR GQAFVIF-C) was cloned into the pB33-eCPX construct [166,167] using restriction enzymes NdeI and XhoI (NEB), downstream of an in-frame myc tag (full sequence given in Section C.4.5and transformed into 5-α competent *E. coli* cells (NEB). The eCPX-β2β3-loop plasmid DNA was purified by miniprep (Omega) and 200 ng were used to electroporate E. coli MC1061 F– cells (Lucigen) in 1 mm electroporation cuvettes (Fisher). Cells were grown in 50 ml LB (Fisher) containing 12.5 μg ml^{-1} chloramphenicol (GoldBio Technology) at 37 °C to an OD600 of 0.5 and induced overnight with 0.1% arabinose at 25 °C. ~5×10^8 cells were pelleted (7300 × *g*) for 5 min at 4 °C, then washed with ice-cold CellGro PBS 1x (Corning). Cells were incubated with 100 nM Cy5-labeled TAR RNA (IDT) (treated as in Section 2.4.2) and 1:10,000-fold diluted FITC-conjugated anti-cMyc antibody (Abcam), cells were incubated, rotating, at 4 °C in 1 ml PBS for 1 h. Cells were pelleted and washed once with ice-cold PBS. RNA-binding (Cy5 fluorescence) and display (FITC fluorescence) were measured using a CyAn ADP flow cytometer (Beckman-Coulter). All flow data were analyzed and plotted using FlowJo 10.3.

5.8.2 Yeast Display

The β2β3 loop peptide was cloned into Aga2 by cutting pCTcon2 with NheI and BamHI without CIP treatment, and gel extracting the cut plasmid. The β2β3 sequence was inserted by ordering primers which, upon annealing, form overhangs complementary to the cut vector, and mixing them at a 50:1 ratio with the cut vector under the conditions of a Quick Ligase Kit (NEB). The sequence is given in Section C.4.4, and is the *full* β2β3 loop, including the β-strands (analogous to *peptide* **1**).

Transformation and induction of β2β3 loop displaying yeast were performed according to the protocol in Section 2.4.2. As in previous yeast display assays, surface display was measured with FITC conjugated anti-myc antibody from Abcam, and RNA binding measured using Cy-5 labelled TAR in concentrations ranging from 1–1000 nM.

5.8.3 Display Assay Results

As can be seen in Figure 5.8A, the β2β3 loop is displayed on the surface of *E. coli*, though not spectacularly. Insofar as it *does* display, it appears to be binding TAR. A comparison of Figure 5.8A to Figure 5.8B indicates that this binding is not due to some innate property of the surface of the bacteria, but instead a result of the β2β3 peptide displayed upon it.

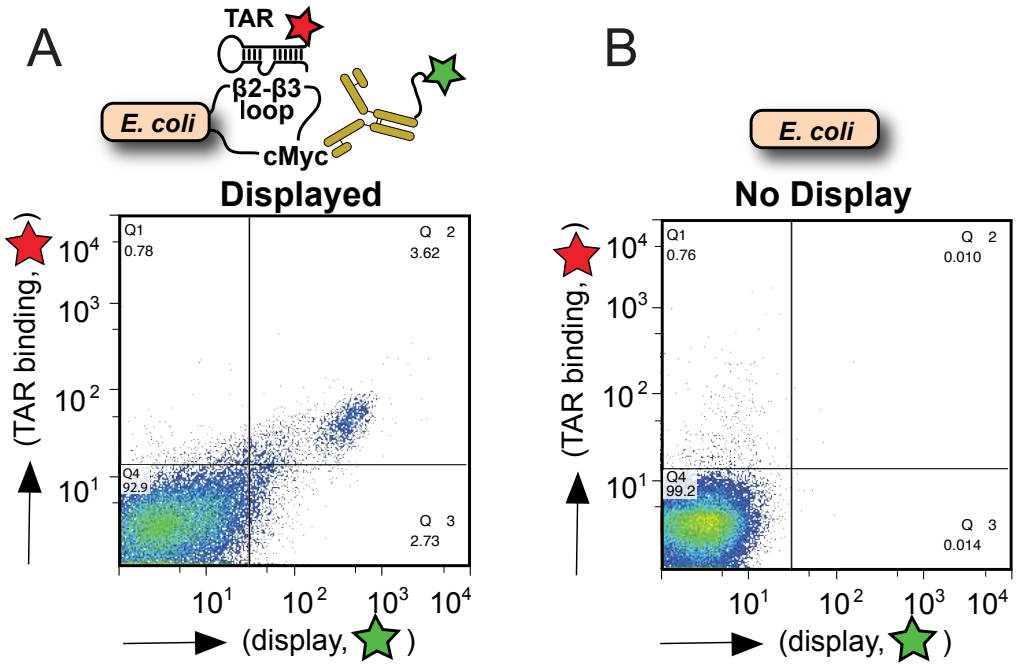


Figure 5.8: Flow Cytometry Analysis of Bacterial Displaying a β2β3 loop Bacterial display of TBP 6.7 β2β3 loop shows binding to TAR that directly correlates with display, indicating that any bacteria which display the β2β3 loop are able to bind TAR, and bacteria that do not express the β2β3 loop do not. From [45].

The results of the analogous yeast display experiment are shown in Figure 5.9A, and indicate that the β2β3 loop displays on yeast better than it displays on bacteria. This is still somewhat poor display, and certainly worse than the naked Aga2 control displays. Figure 5.9B shows that

this displayed loop binds TAR well, and that this improved binding persists even at the extremely low concentration of 1 nM.

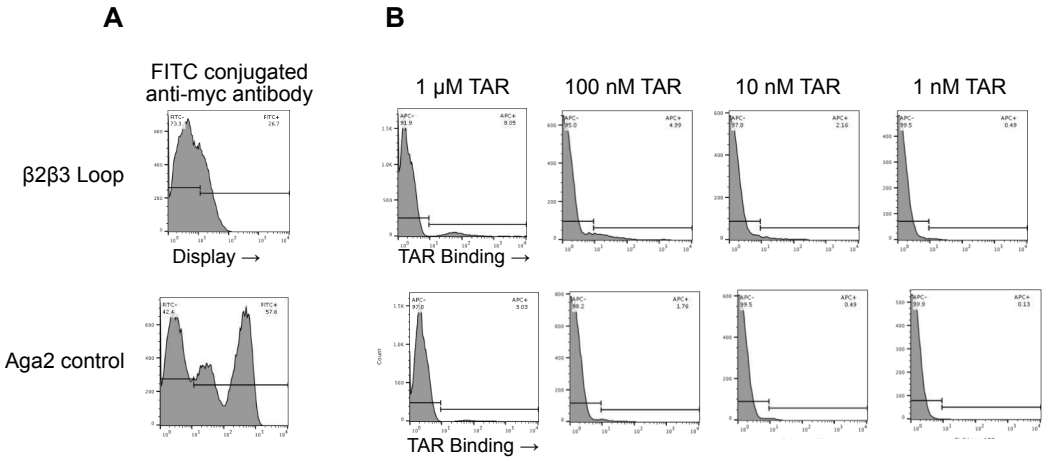


Figure 5.9: Flow Cytometry Analysis of Yeast Displaying a β2β3 loop Display of the β2β3 peptide on yeast results in clear binding of TAR over background across a wide range of conentrations (1–1000 nM). This binding of TAR over background does *not* occur on yeast displaying only the Aga2 protein which the β2β3 loop peptide is fused to (See Figure 2.3.1).

The possibilities of surface display are obvious. Though there is clearly space to thoughtfully engineer variants of peptides based on the TBP 6.7 β2β3 loop, if past (e.g Chapter 2) is prologue, the best chance of success will come in using high-throughput methods to select for novel RNA-binders. As such, though these initial surface display results are *qualified* successes, they are also quite clearly *important*. Along with the SUMO fusions discussed in Section 5.7, these display assays represent the possibility of synthesizing, characterizing, and engineering RNA binding peptides with all the ease of recombinant protein synthesis and all the power of high-throughput screening.

5.9 Conclusions

The work described in this chapter represents a culmination and expansion of the work done in Chapters 2-4. From a basic research perspective, the work done in this chapter represents a distillation of sorts. We began the affinity maturation process with the assumption that opti-

mizing the β2β3 of the U1A RRM would be a necessary component of optimizing a protein of ~100 residues in order to bind TAR. The data here show that, to a reasonable approximation, an optimized β2β3 loop of ~10 residues is *sufficient* for binding TAR.

Furthermore, this simplified peptide is at least somewhat transferrable to expression and selection platforms. The peptide expresses, and demonstrates TAR binding activity across a variety of assays (e.g. ITC, ELISA, Flow Cytometry) and within diverse milieus (e.g. β–turn–β peptide, isolated loop peptide, SUMO fusion, membrane protein fusion). This flexible, minimal functionality indicates that we not only succeeded in finding an extraordinary TAR-binding *protein*, but a possibly transferrable, modular *motif* for binding TAR.

Chapter 6

Conclusions and Future Directions

Here at the end of the story of the development of binders for TAR RNA, let's take a look back at the initial goals of this project, and a look forward at what the next steps might be.

6.1 Project Background and Goals

The work in this thesis was inspired by the recent explosion in understanding of the many roles of cellular RNA, both as a linear sequence and as functional, structured elements. RNA sits in the center of the Central Dogma of Molecular Biology [15], and plays a role in regulating every step from gene regulation [25] to mRNA translation [168] (Figure 1.2). Furthermore, the well-known canonical roles of tRNA and rRNA in facilitating the formation of the peptide bond make up an overwhelming majority of the RNA in a cell (Table 1.1).

Even as the extraordinary variety of RNA function becomes apparent, the fact that such a huge majority of the RNA in a cell is "untouchable" rRNA and tRNA, combined with the limited chemical diversity of RNA (only 4 bases which are of similar chemical character) make targeting a *specific* RNA a major challenge.

There are already a variety of methods for targeting arbitrary single-stranded RNA (most notably modular, engineerable PUF proteins, Section 1.15), but targeting structured RNA remains a challenge. The structured RNA that primarily concerns this thesis is the TAR element of HIV-1, a structured RNA that plays an important role as a miRNA [46], and, through its association with the HIV-1 Tat protein, as a transcriptional activator [40–42].

This project had both a concrete and abstract goal. The concrete goal was difficult, but straightforward: generate a binder for the structured and functional TAR element of HIV using modern protein engineering tools.

The abstract goal was that in so doing we might learn some general information about protein–RNA interactions. This was really an abstract *hope*, since any successes in this realm

would be informed by the progress of the concrete project, not planned from the outset. In both of these goals we succeeded beyond any expectation at the outset. A brief visual summation of the project is shown in Figure 6.1

6.2 Achievement of Project Goals

6.2.1 Develop a Protein-Based Binder of TAR RNA

Affinity Maturation Assumptions

In the concrete goal of finding a binder for TAR RNA, our plan of "semi-design" operated on the assumption that the U1A RRM would serve as a scaffold, and that the *broadened* specificity of the U1A E19S variant could be a starting point for truly *altered* specificity. We hoped that targeted changes to the U1A scaffold could unlock new, specific interactions for a different RNA target, while maintaining the RRM fold that Nature has found so successful at binding a variety of RNA structures [99].

Library Screening

We chose to use the yeast display technique to analyze our library, and randomized the β2β3 loop (for six rounds) and the C-helix (for three rounds) regions of the U1A protein. After these six rounds of selection, we seemed to achieve that goal, with our final round of screening having notable levels of RNA binding even at extremely low concentrations of TAR RNA, and high concentrations of competitor tRNA Figure 2.5. The final library showed excellent convergence to a consensus sequence in the β2β3 loop (Figure 2.8. This represents the first use of yeast display to select for an RNA binding protein for a specific RNA.

Protein Characterization

When we characterized the resulting proteins, by a variety of techniques, we found that our best binder (TBP 6.7) binds TAR with exceptional affinity (K_D = ~500 pM, Figure 3.8), an order of magnitude better than our initial hopes of finding a binder with K_D =~ 10 nM.

In addition to its high affinity, TBP 6.7 is exquisitely *selective* for TAR. There is minimal left-over affinity for U1hpII (the original cognate RNA of U1A), and the ability of TBP 6.7 to bind TAR was reasonably sensitive to mutation (Figure 3.9). To our delight, this protein also seemed to perform well in biochemically relevant competition and transcription assays, inhibiting Tat/TAR-dependent transcription (Figure 3.17).

6.2.2 Advance Understanding of Protein–RNA Binding

Our collaborators at the University of Rochester were able to obtain a co-crystal structure of the TBP 6.7–TAR complex (PDB:6cmn, Figure 4.4). This crystal structure was groundbreaking for a variety of reasons. The primary reason is that it represents the first full crystal structure of the frequently targeted TAR element. In a more general sense, it represents one of a *very* few examples of a crystal structure involving a synthetic protein/RNA interaction. Among the most notable findings are the seemingly dichotomous understandings that the basic structure of TBP 6.7 is extremely similar to that of the U1A protein it is derived from (confirming our "semi-design" hypothesis), while the *mode* of binding is fundamentally altered (TBP 6.7 binds the major groove, rather than the single-stranded loop as in the U1A–U1hpII interaction).

My own work has been focused on creation and characterization of properties. This crystal structure, which elucidates *how* the TAR/TBP 6.7 interaction occurs, rather than simply characterizing the interaction, certainly represents achievement of this abstract goal of advancing understanding protein–RNA bindng. In a broader sense, I hope that the work by myself and my wonderful collaborators will be a source of inspiration and direction for the field. Given the detail and novelty of the TBP 6.7–TAR interaction, I have little doubt in that outcome.

6.2.3 Develop a Peptide Based Binder of TAR RNA

In fact, the first success of this abstract goal of providing new insight into protein/RNA interaction actually comes to fruition within this thesis. The revelation of the importance of the β2β3 loop in the crystal structure inspired the creation and analysis of short peptide derivatives of the

β2β3 loop of this protein. These peptides demonstrated reasonable affinity (K_D = 1.8 μM, Figure 4.7) and specificity for TAR, and also disrupt Tat/TAR-dependent transcription (Figure 5.6).

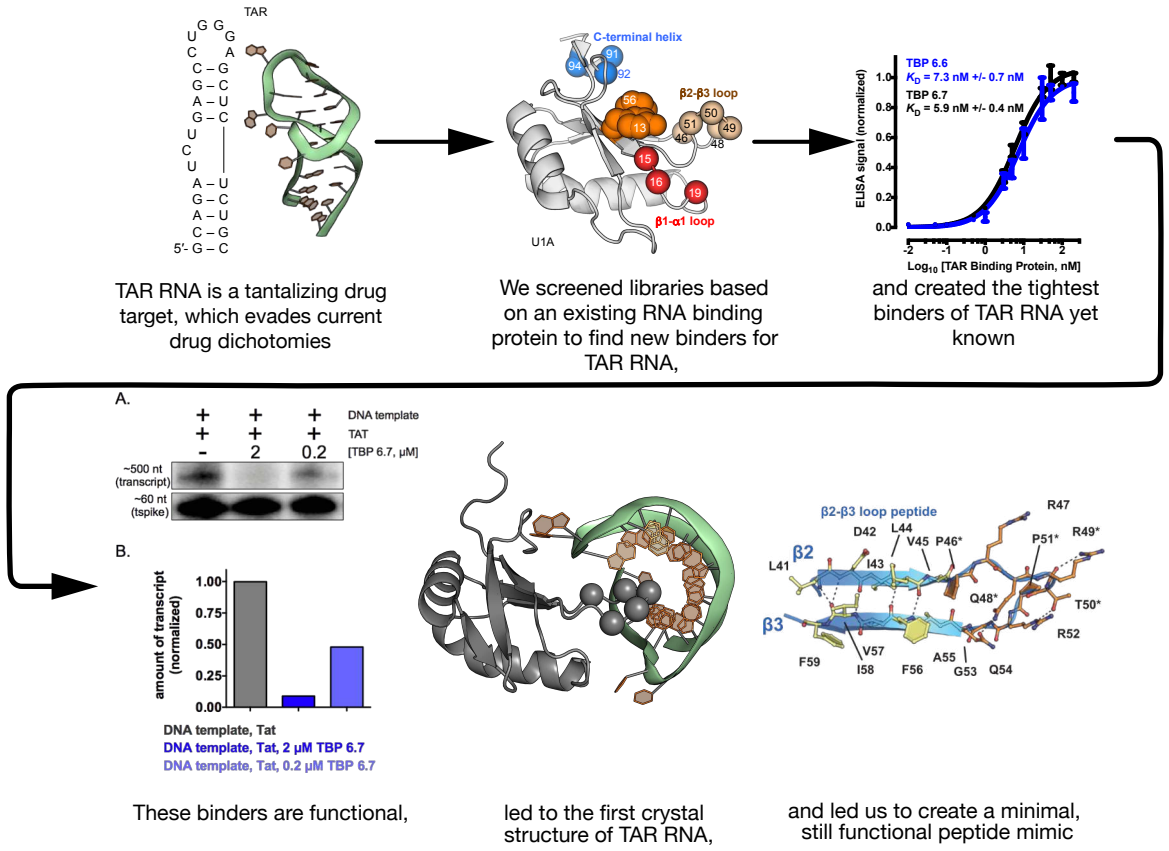


Figure 6.1: Project Summary The included selections from various figures throughout this thesis summarize the milestones and successes along the path to achieving our goal of finding a binder to the HIV-1 TAR element, and the advances made possible by achieving that concrete goal.

6.3 Future Directions

6.3.1 Optimization of Peptide Derivatives of TBP 6.7

Given the general knowledge gained in this work, the future directions are endless, but there are a few clear areas of possibility. The first is the thoughtful refinement of the peptides derived from the β2β3 loop of TBP 6.7. The fact is that the peptides studied in this thesis were extracted from the context of a full-length protein. Though they are already functional in their current

form, there was no particular effort to optimize them for independent existence as isolated peptides, and doing so will be an obvious first step.

The "strand" of the project most within my own skill set and knowledge base is in protein engineering to determine the modularity of these peptides.

β-strand Optimization

Given that *peptide* **1s**, which consists only of the β2β3 *loop* absent the connecting β-strands, is still active via ITC (Section 5.5) and *in vitro* transcription (Section 5.6) the most logical place to start optimizing the peptide would be to β-strands using motifs found in Nature, using canonical amino acids. A good source of inspiration here is the lab of Niels Andersen, which has published many papers detailing the use of aryl-aryl interactions (functionally similar to the π-stacking interactions that stabilize protein–RNA interactions) on opposing strands to make ultra-stable β-turn-β motifs. It is likely possible to combine the loop sequence with optimized β-strands inspired by stable motifs found in thermophilic organisms, see [169] and [170] for an idea of what these strands could consist of.

Some work has already been done in this area by my colleague Patrick Beardslee. The tested peptides were made using β-strands stabilized by cation-π interactions, directly inspired by work from Marcey Waters' lab [171].

Non-canonical Peptides

In addition to the use of optimized peptides based on the canonical amino acids, the β-strands could be optimized using non-canonical amino acids, especially used as capping (such as in [172]). In point of fact, *peptide* **1** and *peptide* **1s**, used extensively in Chapter 5, represent a version of this strategy, since they are forced into a cyclic conformation using a small molecule.

6.3.2 Optimized Surface Display

Finally, any modified peptides should be placed into a position where the same powerful library screening tools we used to select a TAR binding *protein* could be used to select for a

TAR binding *peptide*. This would be the most important application of stabilized β-strands using canonical amino acids, since it would allow more stable variation in the loop structure. The cyclization strategies used to make the synthetic peptides could also be applied to surface display in order to make more rigid structures (see Section B.2 for some initial attempts).

6.3.3 Optimized Recombinant Expression

Any TAR binding peptides discovered through surface display screening would be best analyzed with a simple recombinant platform to stabilize the peptide loop. The method and data shown in Section 5.7 are a first step, but certainly don't represent a comprehensive engineering program, especially given the low ELISA signal. Experimenting with different linker lengths and fusion domains is an obvious next step.

6.4 Progress Toward a Binding Code

The knowledge of the TBP 6.7–TAR co-crystal structure might well be an important step in the dream of building a comprehensive code for binding structured RNA. Though any suggestions in this section is, by definition, speculation, there *is* a general strategy that could be employed, and general hot-spots of interest on the protein.

Going back to basics (Section 1.10), an RNA binder needs to

- Embrace the broad chemical characteristics of RNA
- Accommodate, or guide into accommodation, the general structure of the RNA molecule in question
- Make *specific*, base-dependent interactions to generate selectivity

For a universal structured RNA to exist it would be be necessary that there be some modularity, though what form it takes would be impossible to know *a priori*. My own hypothesis is that there will be a "flow-chart" of sorts, based on the three characteristics discussed above.

At the first level of the flow chart (corresponding to points 1 and 2) would be very general properties such as size, degree of helicity, and degree of screw-axis distortion. For example, a miRNA (with repeated internal loops) would be different than TAR (with a single bulge), and they would both be different than a pseodoknot-containing RNA element. These differences in structure would likely mean different *classes* of RNA binding protein. As such, this work probably means generally less to the goal of binding the SL1 element of the HIV-1 dimerization initiation site (DIS), which is a pseudoknot, than it does to the goal of binding the SL2 element, which is a short stem-loop with a single base bulge (see Figure 3.11 for a close look at this portion of the HIV-1 genome).

6.4.1 Structural Considerations

The second half of point 2, accommodating general structure, could be understood to some degree based on deviations from the *specific* TAR/TBP 6.7 interaction. The best way to determine the beginnings of a code is likely the use of compensatory mutants.

To some degree, this has already been tried. For instance, hp5 and hp6 in Figure 3.9 were designed to make it easier for the loop-adjacent stem residues on TAR to de-hybridize. If the mode of binding for TBP 6.7–TAR was similar to that of U1A–U1hpII (single-stranded binding to the large loop), this would have minimally disrupted, or even improved binding. As it is, these mutations decrease TBP 6.7 binding drastically, but do not abolish it to the degree of the ΔUCU in hp1. This was our first clue that the binding modes were, indeed, different, and we would learn that TBP 6.7 largely contacts the major groove of the 8 paired residues between the loop and the bulge, which means that it is sensitive to changes in these residues.

I would propose learning the extent of this double-stranded binding activity, since it seems based on broad structural shape (the TBP 6.7 β2β3 is generally twisted on its axis compared to the β2β3 loop of wtU1A, and somewhat narrowed), which is seemingly caused by the two introduced proline mutations, especially P51. I assume that a ΔC39ΔG36 TAR would probably not

bind TBP 6.7, but shortening (narrowing?) the β2β3 loop in compensation (I would propose a ΔQ54 mutant) might restore it.

Likewise, *extending* the helical region to the point where the loop no longer impacted TBP 6.7 binding could also be illuminating, and would be the real beginnings of a code for "bulge-helix" regions of RNA. My own guess is that adding 2–3 extra base pairs beyond the C29·G36 region would abolish binding. This binding could potentially be restored to some degree, and the means of doing so would be illustrative. Possible means of restoration could be the addition of a residue to the loop near position 54, or mutating Q54 (which participates in a single H-bond interaction with the backbone of G34) to a residue which would facilitate non-specific interactions lysine (good for cation-anion interactions) or a threonine (able to facilitate H-bonds or π-stacking).

Obviously, there are wild-cards that probably cannot be accounted for in a quest for a reductive system. For TAR, notably, the reduction in binding seen in hp2 and hp4 in Figure 3.9 had less to do with contacts between TAR and TBP 6.7, and more to do with the abolishment of the cross-loop base pair between C30 and G34 which stabilizes the TBP 6.7–TAR interaction. But it could still be possible to get to "good-enough" based on the TBP 6.7–TAR interaction.

6.4.2 Base Interactions

More difficult to conceptualize is adapting the pieces of the TBP 6.7–TAR interaction that correspond to point 3 – specific interactions to base pairs. What is interesting about this set of interactions is that it was *not*, as we had assumed, merely a matter of introducing the correct residues to make the contact, but in *placing* existing residues. Of the three vital arginine residues—R47, R49, and R52—only R49 was introduced via affinity maturation.

As such, it seems to me that it won't simply be a matter of replacing or moving single arginine mutations, but in making and screening a series of small, targeted libraries. As an example, I'll discuss modifying the role of R47, which contacts three separate bases, most notably two H-bonds to G26.

The simplest conceptualization is modifying the binding to a G26C/C39G double mutation. I would make a series of small, easy to screen libraries based on randomizing R47 and two other residues, notably positions 46+48 and 45+49. Randomizing 3 positions results in a small library of $20^3 = 8000$ variations, which is easily screened many times over, and more importantly is less likely to fundamentally alter the binding mode, which should enable step-by-step alteration of binding *interaction* (rather than the wholesale change in binding *mode* represented by the affinity maturation in Chapter 2), which could feasibly become a set of rationally understood binding interactions, and therefore predictably *adaptable*.

A rationally designed binder for the SL2 stem-loop element on the HIV-1 DIS may be far in the future, but a rationally designed binder for a slight variation of the TAR element (for instance, as a result of HIV-1 mutation in response to the pressure exerted by a TAR-binding pharmaceutical) might not be such a pipe dream. Again, the beauty of the TAR binding solution represented by TBP 6.7 was the combination of new binding residues and the altered *placement* of the old ones. A full understanding of how these placements occurred would be a remarkable tool.

6.4.3 Conclusions

Though the field is probably decades from the dream of a simple, modular "code" for structured RNA binding, I believe it *is* possible. The limited chemical and structural diversity of RNA may make targeting a *single* RNA difficult, but it also means that though the array of targets may be vast, it is also finite. There are a limited number of structural motifs, and a limited (and predictable) number of base-pairing interactions. My dearest hope is that the work discussed in this thesis, and the future directions discussed in this chapter, will enable advances in the fundamental understanding of the ways in which proteins interact with RNA.

Bibliography

[1] Mulder, GJ. Sur la composition de quelques substances animales. *Bulletin des Sciences Physiques et Naturelles en Néerlande*, pages 104–124, **1838**

[2] Hartley, H. Origin of the Word 'Protein'. *Nature*, 168(4267):244–244, **1951**

[3] Sumner, JB. The Isolation and Crystallization of the Enzyme Urease. *The Journal of biological chemistry*, 69(2):435–441, **1926**

[4] All Nobel Prizes in Chemistry. URL https://www.nobelprize.org/prizes/lists/all-nobel-prizes-in-chemistry/

[5] Pauling, L, Corey, RB, and Branson, HR. The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain. *Proceedings of the National Academy of Sciences*, 37(4):205–211, **1951**

[6] Milo, R. *Cell biology by the numbers*. Garland Science, Taylor & Francis Group, New York, NY, **2016**

[7] Lodish, H. *Molecular cell biology*. W.H. Freeman, New York, **2008**

[8] Beadle, GW and Tatum, EL. Genetic Control of Biochemical Reactions in Neurospora. 27(11):499–506, **1941**

[9] Avery, OT, Macleod, CM, and McCarty, M. Studies on the Chemical Nature of the Substance Inducing Transformation of Pneumoccal Types. *Journal of Experimental Medicine*, 79(2):137–158, **1944**

[10] Griffiths, A. *An introduction to genetic analysis*. W.H. Freeman, New York, **2000**

[11] Hershey, AD and CHASE, M. Independent functions of viral protein and nucleic acid in growth of bacteriophage. *The Journal of General Physiology*, 36(1):39–56, **1952**

[12] Watson, JD and Crick, FH. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Science*, 171(4356):737–738, **1953**

[13] Cobb, M. 1953: When Genes Became "Information". *Cell*, 153(3):503–506, **2013**

[14] Crick, FH. On protein synthesis. *Symposia of the Society for Experimental Biology*, 12:138–163, **1958**

[15] Crick, FH. Central Dogma of Molecular Biology. *Nature*, 227(5258):561–563, **1970**

[16] Lander, ES. The new genomics: global views of biology. *Science*, 274(5287):536–539, **1996**

[17] MacDonald, ME, Ambrose, CM, Duyao, MP et al. A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell*, 72(6):971–983, **1993**

[18] An, MC, Zhang, N, Scott, G et al. Genetic Correction of Huntington's Disease Phenotypes in Induced Pluripotent Stem Cells. *Cell Stem Cell*, 11(2):253–263, **2012**

[19] Naldini, L. Gene therapy returns to centre stage. *Nature*, 526(7573):351–360, **2015**

[20] Gori, JL, Hsu, PD, Maeder, ML et al. Delivery and Specificity of CRISPR/Cas9 Genome Editing Technologies for Human Gene Therapy. *Human Gene Therapy*, 26(7):443–451, **2015**

[21] Miller, AD. Human gene therapy comes of age. *Nature*, 357(6378):455–460, **1992**

[22] Gibson, DG, Glass, JI, Lartigue, C et al. Creation of a bacterial cell controlled by a chemically synthesized genome. 329(5987):52–56, **2010**

[23] Fica, SM, Tuttle, N, Novak, T et al. RNA catalyses nuclear pre-mRNA splicing. *Nature*, 503(7475):229–234, **2013**

[24] Geisler, S and Coller, J. RNA in unexpected places: long non-coding RNA functions in diverse cellular contexts. *Nature reviews Molecular cell biology*, 14(11):699–712, **2013**

[25] Costa, FF. Non-coding RNAs: Meet thy masters. *BioEssays*, 32(7):599–608, **2010**

[26] Bass, BL and Cech, TR. Specific interaction between the self-splicing RNA of Tetrahymena and its guanosine substrate: implications for biological catalysis by RNA. *Nature*, 308(5962):820–826, **1984**

[27] Altman, S, Baer, MF, Bartkiewicz, M et al. Catalysis by the RNA subunit of RNase P–a minireview. *Gene*, 82(1):63–64, **1989**

[28] Gilbert, W. Origin of life: The RNA world. 319(6055):618–618, **1986**

[29] Scotti, MM and Swanson, MS. RNA mis-splicing in disease. *Nature Reviews Genetics*, 17(1):19–32, **2015**

[30] Warf, MB, Nakamori, M, Matthys, CM et al. Pentamidine reverses the splicing defects associated with myotonic dystrophy. *Proceedings of the National Academy of Sciences of the United States of America*, 106(44):18,551–18,556, **2009**

[31] Wilson, RC and Doudna, JA. Molecular Mechanisms of RNA Interference. *Annual Review of Biophysics*, 42(1):217–239, **2013**

[32] Ha, M and Kim, VN. Regulation of microRNA biogenesis. *Nature reviews Molecular cell biology*, 15(8):509 EP —-524, **2014**

[33] Nussbacher, JK and Yeo, GW. Systematic Discovery of RNA Binding Proteins that Regulate MicroRNA Levels. *Mol Cell*, 69(6):1005–1016.e7, **2018**

[34] Pfeffer, SR, Yang, CH, and Pfeffer, LM. The Role of miR-21 in Cancer. *Drug Development Research*, 76(6):270–277, **2015**

[35] Adli, M. The CRISPR tool kit for genome editing and beyond. *Nature Communications*, 9(1):1–13, **2018**

[36] Jahan, N, Wimmer, E, and Mueller, S. Polypyrimidine Tract Binding Protein-1 (PTB1) Is a Determinant of the Tissue and Host Tropism of a Human Rhinovirus/Poliovirus Chimera PV1(RIPO). *PLOS ONE*, 8(4):1–11, **2013**

[37] Shehu-Xhilaga, M, Crowe, SM, and Mak, J. Maintenance of the Gag/Gag-Pol Ratio Is Important for Human Immunodeficiency Virus Type 1 RNA Dimerization and Viral Infectivity. *Journal of Virology*, 75(4):1834–1841, **2001**

[38] Low, JT, Garcia-Miranda, P, Mouzakis, KD et al. Structure and Dynamics of the HIV-1 Frameshift Element RNA. *Biochemistry*, 53(26):4282–4291, **2014**

[39] Foley, B, Leitner, T, Apetrei, C et al. HIV Sequence Compendium. In *HIV Sequence Compendium*, pages LA–UR–16–25,625. Los Alamos National Laboratory, **2016**

[40] Huthoff, H and Berkhout, B. Mutations in the TAR hairpin affect the equilibrium between alternative conformations of the HIV-1 leader RNA. *Nucleic Acids Res*, 29(12):2594–2600, **2001**

[41] Harrich, D, Ulich, C, and Gaynor, RB. A critical role for the TAR element in promoting efficient human immunodeficiency virus type 1 reverse transcription. *Journal of Virology*, 70(6):4017–4027, **1996**

[42] Feng, S and Holland, EC. HIV-1 Tat trans-activation requires the loop sequence within TAR. *Nature*, 334(6178):165–167, **1988**

[43] Karn, J and Stoltzfus, CM. Transcriptional and posttranscriptional regulation of HIV-1 gene expression. *Cold Spring Harbor Perspectives in Medicine*, 2(2):a006,916–a006,916, **2012**

[44] Peterlin, BM and Price, DH. Controlling the elongation phase of transcription with P-TEFb. *Mol Cell*, 23(3):297–305, **2006**

[45] Belashov, IA, Crawford, DW, Cavender, CE et al. Structure of HIV TAR in complex with a Lab-Evolved RRM provides insight into duplex RNA recognition and synthesis of a constrained peptide that impairs transcription. *Nucleic Acids Res*, 154:766–15, **2018**

[46] Klase, Z, Winograd, R, Davis, J et al. HIV-1 TAR miRNA protects against apoptosis by altering cellular gene expression. *Retrovirology*, 6(1):18, **2009**

[47] Ouellet, DL, Plante, I, Landry, P et al. Identification of functional microRNAs released through asymmetrical processing of HIV-1 TAR element. *Nucleic Acids Res*, 36(7):2353–2365, **2008**

[48] Klase, Z, Kale, P, Winograd, R et al. HIV-1 TAR element is processed by Dicer to yield a viral micro-RNA involved in chromatin remodeling of the viral LTR. *BMC Mol Biol*, 8(1):63, **2007**

[49] Cuevas, JM, Geller, R, Garijo, R et al. Extremely High Mutation Rate of HIV-1 In Vivo. *PLoS biology*, 13(9):e1002,251–19, **2015**

[50] Kawashima, Y, Pfafferott, K, Frater, J et al. Adaptation of HIV-1 to human leukocyte antigen class I. *Nature*, 458(7238):641–645, **2009**

[51] Coffin, J and Swanstrom, R. HIV pathogenesis: dynamics and genetics of viral populations and infected cells. *Cold Spring Harbor Perspectives in Medicine*, 3(1):a012,526, **2013**

[52] AIDS, JUNPoH. Global AIDS Update 2016, **2016**

[53] Dieffenbach, CW and Fauci, AS. Thirty years of HIV and AIDS: future challenges and opportunities. *Ann Intern Med*, 154(11):766–771, **2011**

[54] Stelzer, AC, Frank, AT, Kratz, JD et al. Discovery of selective bioactive small molecules by targeting an RNA dynamic ensemble. *Nature chemical biology*, 7(8):553–559, **2011**

[55] Davidson, A, Patora-Komisarska, K, Robinson, JA et al. Essential structural requirements for specific recognition of HIV TAR RNA by peptide mimetics of Tat protein. *Nucleic Acids Res*, 39(1):248–256, **2011**

[56] Davidson, A, Leeper, TC, Athanassiou, Z et al. Simultaneous recognition of HIV-1 TAR RNA bulge and loop sequences by cyclic peptide mimics of Tat protein. *Proc Natl Acad Sci U S A*, 106(29):11,931–11,936, **2009**

[57] Richter, S, Parolin, C, Gatto, B et al. Inhibition of human immunodeficiency virus type 1 tat-trans-activation-responsive region interaction by an antiviral quinolone derivative. *Antimicrob Agents Chemother*, 48(5):1895–1899, **2004**

[58] Mei, HY, Mack, DP, Galan, AA et al. Discovery of selective, small-molecule inhibitors of RNA complexes–I. The Tat protein/TAR RNA complexes required for HIV-1 transcription. *Bioorganic & medicinal chemistry*, 5(6):1173–1184, **1997**

[59] Searle, MS and Williams, DH. On the stability of nucleic acid structures in solution: enthalpy-entropy compensations, internal rotations and reversibility. *Nucleic Acids Res*, 21(9):2051–2056, **1993**

[60] Leslie, AGW, Arnott, S, Chandrasekaran, R et al. Polymorphism of DNA double helices. *J Mol Biol*, 143(1):49–72, **1980**

[61] Chen, Y and Varani, G. Engineering RNA-binding proteins for biology. *FEBS Journal*, 280(16):3734–3754, **2013**

[62] Nolan, SJ, Shiels, JC, Tuite, JB et al. Recognition of an essential adenine at a protein-RNA interface: Comparison of the contributions of hydrogen bonds and a stacking interaction. *J Am Chem Soc*, 121(38):8951–8952, **1999**

[63] Lalwani, S, Kumar, R, and Gupta, N. Sequence-Structure Alignment Techniques for RNA: A Comprehensive Survey. *Advances in Life Sciences*, 4(1):21–35, **2014**

[64] Sundquist, WI and Klug, A. Telomeric DNA dimerizes by formation of guanine tetrads between hairpin loops. *Nature*, 342(6251):825–829, **1989**

[65] Huang, H, Suslov, NB, Li, NS et al. A G-quadruplex-containing RNA activates fluorescence in a GFP-like fluorophore. *Nature chemical biology*, 10(8):686–691, **2014**

[66] Serganov, A and Nudler, E. A decade of riboswitches. *Cell*, 152(1-2):17–24, **2013**

[67] Dervan, PB and Edelson, BS. Recognition of the DNA minor groove by pyrrole-imidazole polyamides. *Current Opinion in Structural Biology*, 13(3):284–299, **2003**

[68] Dethoff, EA, Hansen, AL, Musselman, C et al. Characterizing complex dynamics in the transactivation response element apical loop and motional correlations with the bulge by NMR, molecular dynamics, and mutagenesis. *Biophys J*, 95(8):3906–3915, **2008**

[69] Dethoff, Ea, Petzold, K, Chugh, J et al. Visualizing transient low-populated structures of RNA. *Nature*, 491(7426):724–728, **2012**

[70] Patwardhan, NN, Ganser, LR, Kapral, GJ et al. Amiloride as a new RNA-binding scaffold with activity against HIV-1 TAR. *Medchemcomm*, 8(5):1022–1036, **2017**

[71] McNaughton, BR, Gareiss, PC, and Miller, BL. Identification of a Selective Small-Molecule Ligand for HIV-1 Frameshift-Inducing Stem-Loop RNA from an 11,325 Member Resin Bound Dynamic Combinatorial Library. *J Am Chem Soc*, 129(37):11,306–11,307, **2007**

[72] Friend, K, Campbell, ZT, Cooke, A et al. A conserved PUF–Ago–eEF1A complex attenuates translation elongation. *Nature structural & molecular biology*, 19(2):176–183, **2012**

[73] Beerli, RR and Barbas, CF. Engineering polydactyl zinc-finger transcription factors. *Nature Biotechnology*, 20(2):135–141, **2002**

[74] Townsend, JA, Wright, DA, Winfrey, RJ et al. High-frequency modification of plant genes using engineered zinc-finger nucleases. *Nature*, 459(7245):442–445, **2009**

[75] Carroll, D. Genome engineering with zinc-finger nucleases. *Genetics*, 188(4):773–782, **2011**

[76] Bogdanove, AJ and Voytas, DF. TAL effectors: customizable proteins for DNA targeting. *Science*, 333(6051):1843–1846, **2011**

[77] Boch, J, Scholze, H, Schornack, S et al. Breaking the code of DNA binding specificity of TAL-type III effectors. *Science*, 326(5959):1509–1512, **2009**

[78] Moscou, MJ and Bogdanove, AJ. A simple cipher governs DNA recognition by TAL effectors. *Science*, 326(5959):1501–1501, **2009**

[79] Deng, D, Yin, P, Yan, C et al. Recognition of methylated DNA by TAL effectors. *Cell Research*, 22(10):1502–1504, **2012**

[80] Deng, D, Yan, C, Pan, X et al. Structural basis for sequence-specific recognition of DNA by TAL effectors. *Science*, 335(6069):720–723, **2012**

[81] Lehmann, R and Nüsslein-Volhard, C. Involvement of the pumilio gene in the transport of an abdominal signal in the Drosophila embryo. 329(6135):167–170, **1987**

[82] Zamore, PD, Bartel, DP, Lehmann, R et al. The PUMILIO-RNA interaction: a single RNA-binding domain monomer recognizes a bipartite target sequence. *Biochemistry*, 38(2):596–604, **1999**

[83] Wang, X, McLachlan, J, Zamore, PD et al. Modular recognition of RNA by a human pumilio-homology domain. *Cell*, 110(4):501–512, **2002**

[84] Wang, X, Zamore, PD, and Hall, TM. Crystal structure of a Pumilio homology domain. *Mol Cell*, 7(4):855–865, **2001**

[85] Edwards, TA, Pyle, SE, Wharton, RP et al. Structure of Pumilio reveals similarity between RNA and peptide binding motifs. *Cell*, 105(2):281–289, **2001**

[86] Cheong, CG and Hall, TMT. Engineering RNA sequence specificity of Pumilio repeats. *Proceedings of the National Academy of Sciences*, 103(37):13,635–13,639, **2006**

[87] Johansson, HE, Liljas, L, and Uhlenbeck, OC. RNA Recognition by the MS2 Phage Coat Protein. *Seminars in Virology*, 8(3):176–185, **1997**

[88] SenGupta, DJ, Zhang, B, Kraemer, B et al. A three-hybrid system to detect RNA-protein interactions in vivo. *Proceedings of the National Academy of Sciences*, 93(16):8496–8501, **1996**

[89] Dong, S, Wang, Y, Cassidy-Amstutz, C et al. Specific and modular binding code for cytosine recognition in Pumilio/FBF (PUF) RNA-binding domains. *The Journal of biological chemistry*, 286(30):26,732–26,742, **2011**

[90] Filipovska, A, Razif, MFM, Nygård, KKA et al. A universal code for RNA recognition by PUF proteins. *Nature chemical biology*, 7(7):425–427, **2011**

[91] Adamala, KP, Martin-Alarcon, DA, and Boyden, ES. Programmable RNA-binding protein composed of repeats of a single modular unit. *Proceedings of the National Academy of Sciences of the United States of America*, 113(19):E2579–88, **2016**

[92] Zhao, YY, Mao, MW, Zhang, WJ et al. Expanding RNA binding specificity and affinity of engineered PUF domains. *Nucleic acids research*, 46(9):4771–4782, **2018**

[93] Miranda, RG, McDermott, JJ, and Barkan, A. RNA-binding specificity landscapes of designer pentatricopeptide repeat proteins elucidate principles of PPR–RNA interactions. *Nucleic Acids Res*, 46(5):2613–2623, **2017**

[94] Klug, A. The Discovery of Zinc Fingers and Their Applications in Gene Regulation and Genome Manipulation. *Annual review of biochemistry*, 79(1):213–231, **2010**

[95] Mackay, JP, Font, J, and Segal, DJ. The prospects for designer single-stranded RNA-binding proteins. *Nature Publishing Group*, 18(3):256–261, **2011**

[96] Lunde, BM, Moore, C, and Varani, G. RNA-binding proteins: modular design for efficient function. *Nature reviews Molecular cell biology*, 8(6):479–490, **2007**

[97] Clery, A, Blatter, M, and Allain, FH. RNA recognition motifs: boring? Not quite. *Current Opinion in Structural Biology*, 18(3):290–298, **2008**

[98] Maris, C, Dominguez, C, and Allain, FHT. The RNA recognition motif, a plastic RNA-binding platform to regulate post-transcriptional gene expression. *The FEBS journal*, 272(9):2118–2131, **2005**

[99] Muto, Y and Yokoyama, S. Structural insight into RNA recognition motifs: Versatile molecular Lego building blocks for biological systems. *Wiley Interdisciplinary Reviews: RNA*, 3(2):229–246, **2012**

[100] Allain, FH, Gubser, CC, Howe, PW et al. Specificity of ribonucleoprotein interaction determined by RNA folding during complex formulation. *Nature*, 380(6575):646–650, **1996**

[101] Law, MJ, Rice, AJ, Lin, P et al. The role of RNA structure in the interaction of U1A protein with U1 hairpin II RNA. *RNA*, 12(7):1168–1178, **2006**

[102] Oubridge, C, Ito, N, Evans, PR et al. Crystal structure at 1.92 A resolution of the RNA-binding domain of the U1A spliceosomal protein complexed with an RNA hairpin. *Nature*, 372(6505):432–438, **1994**

[103] Katsamba, PS, Myszka, DG, and Laird-Offringa, IA. Two functionally distinct steps mediate high affinity binding of U1A protein to U1 hairpin II RNA. *The Journal of biological chemistry*, 276(24):21,476–21,481, **2001**

[104] Law, MJ. The role of positively charged amino acids and electrostatic interactions in the complex of U1A protein and U1 hairpin II RNA. *Nucleic Acids Res*, 34(1):275–285, **2006**

[105] Blakeley, BD and McNaughton, BR. Synthetic RNA recognition motifs that selectively recognize HIV-1 trans-activation response element hairpin RNA. *ACS Chemical Biology*, 9(6):1320–1329, **2014**

[106] Katsamba, PS, Bayramyan, M, Haworth, IS et al. Complex role of the beta 2-beta 3 loop in the interaction of U1A with U1 hairpin II RNA. *The Journal of biological chemistry*, 277(36):33,267–33,274, **2002**

[107] Blakeley, BD, Shattuck, J, Coates, MB et al. Analysis of protein-RNA complexes involving a RNA recognition motif engineered to bind hairpins with seven- and eight-nucleotide loops. *Biochemistry*, 52(28):4745–4747, **2013**

[108] Laird-Offringa, IA and Belasco, JG. Analysis of RNA-binding proteins by in vitro genetic selection: identification of an amino acid residue important for locking U1A onto its RNA target. *Proceedings of the National Academy of Sciences*, 92(25):11,859–11,863, **1995**

[109] Crawford, DW, Blakeley, BD, Chen, PH et al. An Evolved RNA Recognition Motif That Suppresses HIV-1 Tat/TAR-Dependent Transcription. *ACS Chemical Biology*, 11(8):2206–2215, **2016**

[110] Law, MJ, Lee, DS, Lee, CS et al. The role of the C-terminal helix of U1A protein in the interaction with U1hpII RNA. *Nucleic Acids Res*, 41(14):7092–7100, **2013**

[111] Wong, TS, Roccatano, D, and Schwaneberg, U. Steering directed protein evolution: strategies to manage combinatorial complexity of mutant libraries. *Environ Microbiol*, 9(11):2645–2659, **2007**

[112] Wong, TS, Roccatano, D, Zacharias, M et al. A statistical analysis of random mutagenesis methods used for directed protein evolution. *J Mol Biol*, 355(4):858–871, **2006**

[113] Orcutt, KD and Wittrup, KD. Yeast Display and Selections. pages 207–233. Springer Berlin Heidelberg, Berlin, Heidelberg, **2010**

[114] Blakeley, BD. *Expanding the Role of Proteins in Basic Research and Drug Discovery: From Protein-Protein Interactions to RNA Recognition*. Ph.D. thesis, Colorado State University, **2014**

[115] Parker, JS, Roe, SM, and Barford, D. Structural insights into mRNA recognition from a PIWI domain-siRNA guide complex. *Nature*, 434(7033):663–666, **2005**

[116] Yang, S and Temin, HM. A double hairpin structure is necessary for the efficient encapsidation of spleen necrosis virus retroviral RNA. *The EMBO journal*, 13(3):713–726, **1994**

[117] Pandey, S, Agarwala, P, Jayaraj, GG et al. RNA stem-loop to G-quadruplex equilibrium controls mature miRNA production inside the cell. 54(48):7067–7078, **2015**

[118] Tonge, PJ. Drug–Target Kinetics in Drug Discovery. *ACS Chemical Neuroscience*, 9(1):29–39, **2018**

[119] Berkhout, B and van Wamel, JL. The leader of the HIV-1 RNA genome forms a compactly folded tertiary structure. *RNA*, 6(2):282–295, **2000**

[120] Clever, J, Sassetti, C, and Parslow, TG. RNA secondary structure and binding sites for gag gene products in the 5' packaging signal of human immunodeficiency virus type 1. *Journal of Virology*, 69(4):2101–2109, **1995**

[121] Stephenson, JD, Li, H, Kenyon, JC et al. Three-Dimensional RNA Structure of the Major HIV-1 Packaging Signal Region. *Structure*, 21(6):951–962, **2013**

[122] Lu, K, Heng, X, Garyu, L et al. NMR detection of structures in the HIV-1 5'-leader RNA that regulate genome packaging. *Science*, 334(6053):242–245, **2011**

[123] Kenyon, JC, Prestwood, LJ, Le Grice, SFJ et al. In-gel probing of individual RNA conformers within a mixed population reveals a dimerization structural switch in the HIV-1 leader. *Nucleic Acids Res*, 41(18):e174–e174, **2013**

[124] Keane, SC, Heng, X, Lu, K et al. RNA structure. Structure of the HIV-1 RNA packaging signal. *Science*, 348(6237):917–921, **2015**

[125] Lusvarghi, S, Sztuba-Solinska, J, Purzycka, KJ et al. RNA Secondary Structure Prediction Using High-throughput SHAPE. *Journal of Visualized Experiments : JoVE*, (75):50,243, **2013**

[126] Sztuba-Solinska, J, Shenoy, SR, Gareiss, P et al. Identification of Biologically Active, HIV TAR RNA-Binding Small Molecules Using Small Molecule Microarrays. *J Am Chem Soc*, 136(23):8402–8410, **2014**

[127] Jeang, KT, Xiao, H, and Rich, EA. Multifaceted Activities of the HIV-1 Transactivator of Transcription, Tat. *The Journal of biological chemistry*, 274(41):28,837–28,840, **1999**

[128] Kamine, J, Loewenstein, P, and Green, M. Mapping of HIV-1 Tat protein sequences required for binding to Tar RNA. *Virology*, 182(2):570–577, **1991**

[129] Sczepanski, JT and Joyce, GF. Binding of a Structured d-RNA Molecule by an l-RNA Aptamer. 135(36):13,290–13,293, **2013**

[130] Marciniak, RA, Calnan, BJ, Frankel, aD et al. HIV-1 Tat protein trans-activates transcription in vitro. *Cell*, 63(4):791–802, **1990**

[131] Arzumanov, A, Walsh, AP, Liu, X et al. Oligonucleotide analogue interference with the HIV-1 Tat protein-TAR RNA interaction. *Nucleosides Nucleotides Nucleic Acids*, 20(4-7):471–480, **2001**

[132] Oubridge, C, Ito, N, Teo, CH et al. Crystallisation of RNA-protein complexes. II. The application of protein engineering for crystallisation of the U1A protein-RNA complex. *J Mol Biol*, 249(2):409–423, **1995**

[133] Kapust, RB, Tozser, J, Fox, JD et al. Tobacco etch virus protease: mechanism of autolysis and rational design of stable mutants with wild-type catalytic proficiency. *Protein Eng*, 14(12):993–1000, **2001**

[134] Adams, PD, Afonine, PV, Bunkoczi, G et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr D Biol Crystallogr*, 66(Pt 2):213–221, **2010**

[135] McCoy, AJ, Grosse-Kunstleve, RW, Adams, PD et al. Phaser crystallographic software. *J Appl Crystallogr*, 40(Pt 4):658–674, **2007**

[136] Emsley, P, Lohkamp, B, Scott, WG et al. Features and development of Coot. *Acta Crystallogr D Biol Crystallogr*, 66(Pt 4):486–501, **2010**

[137] Winn, MD, Ballard, CC, Cowtan, KD et al. Overview of the CCP4 suite and current developments. *Acta Crystallogr D Biol Crystallogr*, 67(Pt 4):235–242, **2011**

[138] Lawrence, MC and Colman, PM. Shape complementarity at protein/protein interfaces. *J Mol Biol*, 234(4):946–950, **1993**

[139] Terwilliger, TC, Grosse-Kunstleve, RW, Afonine, PV et al. Iterative-build OMIT maps: map improvement by iterative model building and refinement without model bias. *Acta Crystallogr D Biol Crystallogr*, 64(Pt 5):515–524, **2008**

[140] Hodel, A, Kim, SH, and Brunger, AT. Model bias in macromolecular crystal structures. *Acta Crystallographica Section A*, 48(6):851–858, **1992**

[141] Brunger, AT, Adams, PD, Clore, GM et al. Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr D Biol Crystallogr*, 54(Pt 5):905–921, **1998**

[142] Case, DA, Cerutti, DS, Cheatham III, TE et al. AMBER 2017. page AMBER, **2017**

[143] Izadi, S, Anandakrishnan, R, and Onufriev, AV. Building Water Models: A Different Approach. *J Phys Chem Lett*, 5(21):3863–3871, **2014**

[144] Maier, JA, Martinez, C, Kasavajhala, K et al. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J Chem Theory Comput*, 11(8):3696–3713, **2015**

[145] Perez, A, Marchan, I, Svozil, D et al. Refinement of the AMBER force field for nucleic acids: improving the description of alpha/gamma conformers. *Biophys J*, 92(11):3817–3829, **2007**

[146] Zgarbova, M, Otyepka, M, Sponer, J et al. Refinement of the Cornell et al. Nucleic Acids Force Field Based on Reference Quantum Chemical Calculations of Glycosidic Torsion Profiles. *J Chem Theory Comput*, 7(9):2886–2902, **2011**

[147] Romo, TD, Leioatts, N, and Grossfield, A. Lightweight object oriented structure analysis: tools for building tools to analyze molecular dynamics simulations. *J Comput Chem*, 35(32):2305–2318, **2014**

[148] Puglisi, JD, Tan, R, Calnan, BJ et al. Conformation of the TAR RNA-arginine complex by NMR spectroscopy. *Science*, 257(5066):76–80, **1992**

[149] Aboul-ela, F, Karn, J, and Varani, G. The structure of the human immunodeficiency virus type-1 TAR RNA reveals principles of RNA recognition by Tat protein. *J Mol Biol*, 253(2):313–332, **1995**

[150] Structure of HIV-1 TAR RNA in the absence of ligands reveals a novel conformation of the trinucleotide bulge. *Nucleic acids research*, 24(20):3974–3981, **1996**

[151] Tao, J, Chen, L, and Frankel, aD. Dissection of the proposed base triple in human immunodeficiency virus TAR RNA indicates the importance of the Hoogsteen interaction. *Biochemistry*, 36(12):3491–3495, **1997**

[152] Long, KS and Crothers, DM. Characterization of the solution conformations of unbound and Tat peptide-bound forms of HIV-1 TAR RNA. *Biochemistry*, 38(31):10,059–10,069, **1999**

[153] Michnicka, MJ, Harper, JW, and King, GC. Selective isotopic enrichment of synthetic RNA: application to the HIV-1 TAR element. *Biochemistry*, 32(2):395–400, **1993**

[154] Jaeger, JA and Tinoco, I. An NMR study of the HIV-1 TAR element hairpin. *Biochemistry*, 32(46):12,522–12,530, **1993**

[155] Borkar, AN, Bardaro, MFJ, Camilloni, C et al. Structure of a low-population binding intermediate in protein-RNA recognition. *Proceedings of the National Academy of Sciences of the United States of America*, 113(26):7171–7176, **2016**

[156] Kulinski, T, Olejniczak, M, Huthoff, H et al. The apical loop of the HIV-1 TAR RNA hairpin is stabilized by a cross-loop base pair. *The Journal of biological chemistry*, 278(40):38,892–38,901, **2003**

[157] Richter, S, Cao, H, and Rana, TM. Specific HIV-1 TAR RNA loop sequence and functional groups are required for human cyclin T1-Tat-TAR ternary complex formation. *Biochemistry*, 41(20):6391–6397, **2002**

[158] Allain, FH, Howe, PW, Neuhaus, D et al. Structural basis of the RNA-binding specificity of human U1A protein. *The EMBO journal*, 16(18):5764–5772, **1997**

[159] Jessen, TH, Oubridge, C, Teo, CH et al. Identification of molecular contacts between the U1 A small nuclear ribonucleoprotein and U1 RNA. *The EMBO journal*, 10(11):3447–3456, **1991**

[160] Lalonde, MS, Lobritz, MA, Ratcliff, A et al. Inhibition of both HIV-1 reverse transcription and gene expression by a cyclic peptide that binds the Tat-transactivating response element (TAR) RNA. *PLoS Pathog*, 7(5):e1002,038, **2011**

[161] Guan, L and Disney, MD. Recent advances in developing small molecules targeting RNA. *ACS Chemical Biology*, 7(1):73–86, **2012**

[162] Zhan, P, Pannecouque, C, De Clercq, E et al. Anti-HIV Drug Discovery and Development: Current Innovations and Future Trends. *J Med Chem*, 59(7):2849–2878, **2016**

[163] Mousseau, G, Mediouni, S, and Valente, ST. Targeting HIV transcription: the quest for a functional cure. *Transacting functions of human retroviruses*, 389(Chapter 435):121–145, **2015**

[164] Mijalis, AJ, Thomas, DA, Simon, MD et al. A fully automated flow-based approach for accelerated peptide synthesis. *Nature chemical biology*, 13(5):464–466, **2017**

[165] Bradrick, TD and Marino, JP. Ligand-induced changes in 2-aminopurine fluorescence as a probe for small molecule binding to HIV-1 TAR RNA. *RNA*, 10(9):1459–1468, **2004**

[166] Rice, JJ, Schohn, A, Bessette, PH et al. Bacterial display using circularly permuted outer membrane protein OmpX yields high affinity peptide ligands. *Protein Sci*, 15(4):825–836, **2006**

[167] Rice, JJ and Daugherty, PS. Directed evolution of a biterminal bacterial display scaffold enhances the display of diverse peptides. *Protein Engineering Design and Selection*, 21(7):435–442, **2008**

[168] Riemondy, K, Hoefert, JE, and Yi, R. Not miR-ly micromanagers: the functions and regulatory networks of microRNAs in mammalian skin. *Wiley Interdisciplinary Reviews: RNA*, 5(6):849–865, **2014**

[169] Anderson, JM, Kier, BL, Jurban, B et al. Aryl-aryl interactions in designed peptide folds: Spectroscopic characteristics and optimal placement for structure stabilization. *Biopolymers*, 105(6):337–356, **2016**

[170] Anderson, JM, Shcherbakov, AA, Kier, BL et al. Optimization of a β-sheet-cap for long loop closure. *Biopolymers*, 107(3):e22,995–11, **2016**

[171] Riemen, AJ and Waters, ML. Design of highly stabilized beta-hairpin peptides through cation-pi interactions of lysine and n-methyllysine with an aromatic pocket. *Biochemistry*, 48(7):1525–1531, **2009**

[172] Kier, BL, Shu, I, Eidenschink, LA et al. Stabilizing capping motif for beta-hairpins and sheets. *Proceedings of the National Academy of Sciences of the United States of America*, 107(23):10,466–10,471, **2010**

[173] Beaudenon, S and Huibregtse, JM. HPV E6, E6AP and cervical cancer. *BMC Biochemistry*, 9(Suppl 1):S4–7, **2008**

[174] Gillison, ML, Koch, WM, Capone, RB et al. Evidence for a causal association between human papillomavirus and a subset of head and neck cancers. *Journal of the National Cancer Institute*, 92(9):709–720, **2000**

[175] Tran, N, Rose, BR, and O'Brien, CJ. Role of human papillomavirus in the etiology of head and neck cancer. *Head & Neck*, 29(1):64–70, **2006**

[176] Hall, MT, Simms, KT, Lew, JB et al. The projected timeframe until cervical cancer elimination in Australia: a modelling study. *The Lancet Public Health*, **2018**

[177] Song, S, Pitot, HC, and Lambert, PF. The human papillomavirus type 16 E6 gene alone is sufficient to induce carcinomas in transgenic animals. *Journal of Virology*, 73(7):5887–5893, **1999**

[178] Talis, AL, Huibregtse, JM, and Howley, PM. The role of E6AP in the regulation of p53 protein levels in human papillomavirus (HPV)-positive and HPV-negative cells. *The Journal of biological chemistry*, 273(11):6439–6445, **1998**

[179] Martinez-Zapien, D, Ruiz, FX, Poirson, J et al. Structure of the E6/E6AP/p53 complex required for HPV-mediated degradation of p53. 529(7587):541–545, **2016**

[180] Blackwell, HE and Grubbs, RH. Highly Efficient Synthesis of Covalently Cross-Linked Peptide Helices by Ring-Closing Metathesis. *Angewandte Chemie (International ed in English)*, 37(23):3281–3284, **1998**

[181] Walker, SN, Tennyson, RL, Chapman, AM et al. GLUE That Sticks to HIV: A Helix-Grafted GLUE Protein That Selectively Binds the HIV gp41 N-Terminal Helical Region. *ChemBioChem*, 16(Figure 1):219–222, **2015**

[182] Tennyson, RL, Walker, SN, Ikeda, T et al. Helix-Grafted Pleckstrin Homology Domains Suppress HIV-1 Infection of CD4-Positive Cells. *ChemBioChem*, 17(20):1945–1950, **2016**

[183] Walker, SN, Tennyson, RL, Ikeda, T et al. Evaluation of sequence variability in HIV-1 gp41 C-peptide helix-grafted proteins. *Bioorganic & medicinal chemistry*, 26(6):1220–1224, **2018**

[184] Robinson, H, Gao, YG, McCrary, BS et al. The hyperthermophile chromosomal protein Sac7d sharply kinks DNA. *Nature*, 392(6672):202–205, **1998**

[185] Correa, A, Pacheco, S, Mechaly, AE et al. Potent and Specific Inhibition of Glycosidases by Small Artificial Binding Proteins (Affitins). *PLOS ONE*, 9(5):e97,438–12, **2014**

[186] Sidi, AOMO, Babah, KO, Brimer, N et al. Strategies for bacterial expression of protein-peptide complexes: application to solubilization of papillomavirus E6. *Protein expression and purification*, 80(1):8–16, **2011**

[187] Zanier, K, Charbonnier, S, Sidi, AOMO et al. Structural basis for hijacking of cellular LxxLL motifs by papillomavirus E6 oncoproteins. *339*(6120):694–698, **2013**

[188] Pédelacq, JD, Cabantous, S, Tran, T et al. Engineering and characterization of a superfolder green fluorescent protein. *Nature Biotechnology*, 24(1):79–88, **2005**

[189] Behar, G, Bellinzoni, M, Maillasson, M et al. Tolerance of the archaeal Sac7d scaffold protein to alternative library designs: characterization of anti-immunoglobulin G Affitins. *Protein Engineering Design and Selection*, 26(4):267–275, **2013**

[190] Zanier, K, Stutz, C, Kintscher, S et al. The E6AP Binding Pocket of the HPV16 E6 Oncoprotein Provides a Docking Site for a Small Inhibitory Peptide Unrelated to E6AP, Indicating Druggability of E6. 9(11):e112,514–13, **2014**

[191] Browning, C, Hilfinger, JM, Rainier, S et al. The sequence and structure of the 3' arm of the first stem-loop of the human immunodeficiency virus type 2 trans-activation responsive region mediate Tat-2 transactivation. *Journal of Virology*, 71(10):8048–8055, **1997**

[192] Kalhor-Monfared, S, Jafari, MR, Patterson, JT et al. Rapid biocompatible macrocyclization of peptides with decafluoro-diphenylsulfone. *Chemical Science*, 7(6):3785–3790, **2016**

[193] Magliery, TJ, Wilson, CG, Pan, W et al. Detecting protein-protein interactions with a green fluorescent protein fragment reassembly trap: scope and mechanism. 127(1):146–157, **2005**

[194] Kalderon, D, Roberts, BL, Richardson, WD et al. A short amino acid sequence able to specify nuclear location. *Cell*, 39(3 Pt 2):499–509, **1984**

[195] Bruce, VJ and McNaughton, BR. Evaluation of Nanobody Conjugates and Protein Fusions as Bioanalytical Reagents. *Analytical Chemistry*, 89(7):3819–3823, **2017**

[196] Braun, MB, Traenkle, B, Koch, PA et al. Peptides in headlock–a novel high-affinity and versatile peptide-binding nanobody for proteomics and microscopy. *Sci Rep*, 6(1):19,211, **2016**

[197] Ni, TW, Staicu, LC, Nemeth, RS et al. Progress toward clonable inorganic nanoparticles. *Nanoscale*, 7(41):17,320–17,327, **2015**

[198] Erhart, D, Zimmermann, M, Jacques, O et al. Chemical Development of Intracellular Protein Heterodimerizers. *Chemistry and Biology*, 20(4):549–557, **2013**

Appendix A

Helical Grafting of E6AP

A.1 Background

A.1.1 Significance

One of the more interesting aspects of viral infection is the capacity for the infection to lead to secondary side-effects, especially cancer. One of the more common examples of this phenomenon is the causal link between infection with Human Papillomavirus (HPV) and cervical cancer; nearly all (99.7%) of the 470,000 yearly cases of cervical cancer are caused by HPV infection, but HPV infection is also associated with ~¼ of other mucosal cancers such as those of the mouth, tonsils, and throat [173–175]. There is now a vaccine for the most dangerous forms of HPV (and it is so effective that Australia is set to all-but eliminate new cases of cervical cancer by 2020 [176]), but there will still be ongoing need for treatment, and the problem is of general biochemical interest.

The two proteins most frequently involved in HPV-associated cervical cancer are the cooperative viral oncoproteins E6 and E7, which both affect tumor suppression pathways. E6 affects the p53 tumor suppressor, and E7 affects the retinoblastoma (pRb) suppressor. The research in this section focuses solely on E6.

A.1.2 The E6/E6AP/p53 Ternary Complex

Though the viral protein E6 has at least some p53-independent oncogenic activity [177], its primary means of oncogenesis is by association with p53 by way of an association with E6 associated protein (E6AP) [178], which is also known as Ubiquitin-protein ligase E3A (UBE3A). As the name indicates, E6AP is a ubiquitin ligase, and part of the proteasome. Roughly speaking, E6 is able to bind p53 when it is also binding E6AP, but is not able to bind p53 when it is *not* binding E6AP.

For a detailed expansion of this interaction see [173], which is an excellent review of the role the various regions of E6 and E6AP play in HPV-associated cancers.

For the purposes of this research, the only region of E6AP that matters is the the short epitope that which binds E6. This epitope is a region ~15 residues in length which is largely α-helical in character. E6 binds this short epitope on E6AP independently of p53, and neither E6 nor E6AP bind p53 independently, but the E6/E6AP complex is able to bind p53 to form the E6/E6AP/p53 ternary complex. For the remainder of this chapter, the abbreviation "E6AP" will be assumed to refer to the E6 binding epitope of E6AP, rather than the full protein. A crystal structure of this interaction, the formal logic of the E6/E6AP/p53 ternary complex and the consequences of this interaction are illustrated in simple terms in Figure A.1.

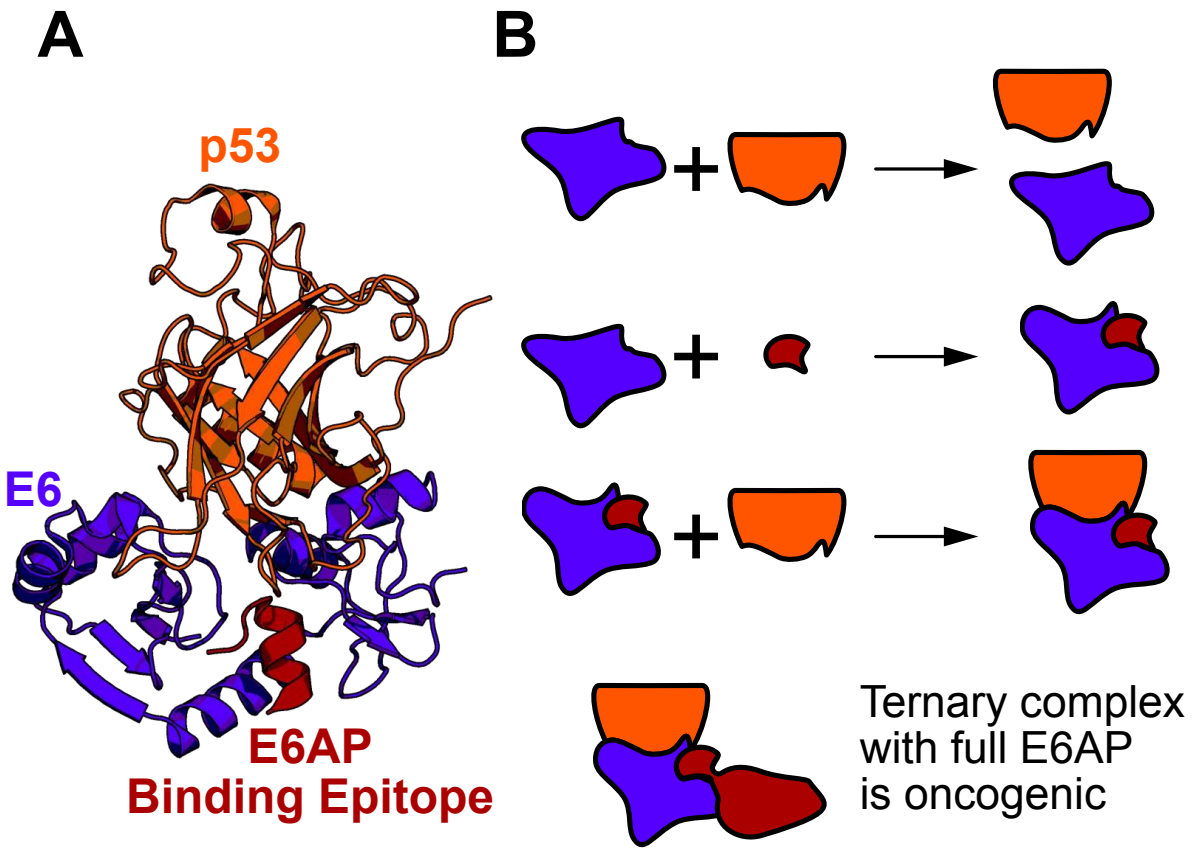


Figure A.1: Crystal Structure and Cartoon of E6/E6AP/p53 Complex The crystal structure in A ([179], PDB: 4xr8) shows the E6/E6AP/p53 ternary complex. The cartoon in **B** is a reductive explanation of the biochemical consequences. In brief, E6 can independently bind E6AP, and the E6/E6AP/p53 ternary complex occurs if and only if E6 is binding E6AP. The formation of the ternary complex trigger the oncogenic pathway. Note that this is true only in a *biochemical* setting when the full-length E6AP protein is present.

A.1.3 The E6/E6AP Binding Interaction

The HPV E6 protein is capable of binding human E6AP protein (UBE3A) even without the presence of p53. This interaction occurs on a leucine-rich α-helical region located on positions 403–414, with the sequence N-LTLQELLGEER-C). Both the leucine-rich helix and the unstructured 3-residue C-terminal tail play an important role in binding to the E6 binding cleft. As such, both hydrophobic and electrostatic contacts are important for this binding interaction

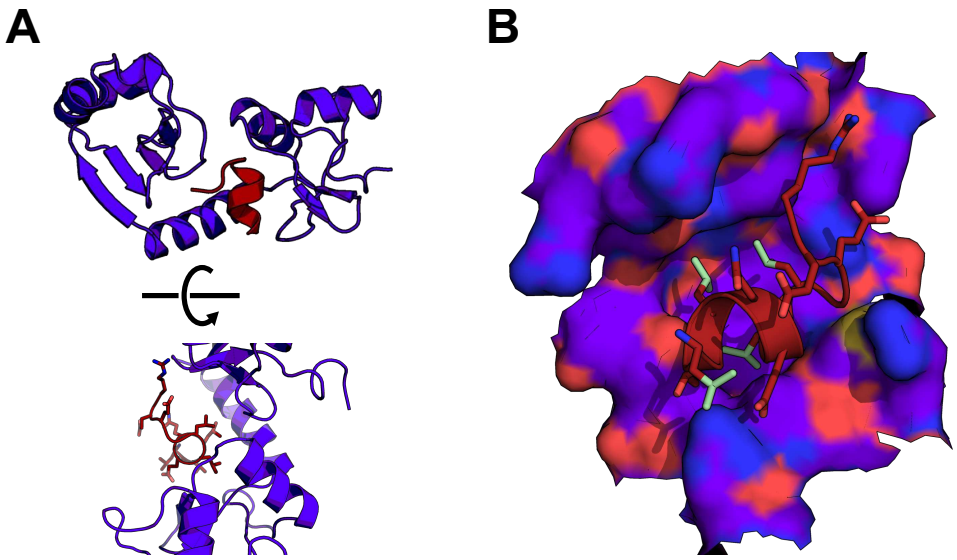


Figure A.2: Detail of the E6/E6AP Binding Interaction The cartoons in **A** show the α-helical region of E6AP binding to its cleft in E6. The surface view in **B** shows the same, with leucines highlighted in teal. Red surface = negative charge, blue surface = positive charge. Note the tight pocket around Arg414 of E6AP.

A.1.4 Research Goals

Long-term Goal

The *ultimate* goal of this research was to create a molecule which could disrupt the formation of this ternary complex. The assumption was that such a molecule might take the form of a short peptide, possibly modified to stabilize a helix, with methods conceptually similar to classic peptide stapling [180]. The peptide would be based on the E6AP binding epitope. Given that a free E6AP binding epitope could plausibly disrupt the interaction between E6 and full-length E6AP, a direct analogue would be acceptable, but ideally the inhibiting peptide would be engineered

thoughtfully, and through high-throughput screening, to have greater affinity for E6 than the wild-type E6AP.

Short-term Goal

The inherent challenge of that stated aim is the difficulty of screening a large number of peptide variants, *especially* when those peptide variants are likely to require modification to fold into the necessary shape (in this case, an α-helix). As such, my proximal goal was to build and characterize a platform which would enable rapid study of the E6/E6AP interaction, and high-throughput screening of same. Additionally, the platform must be expandable to include the study and screening of the interactions of the ternary system—the interactions between E6, E6AP (or engineered variant of E6AP), and p53.

A.2 Helical Grafting Strategy

A.2.1 Helical Stabilization

The primary engineering challenge is the creation of a molecule which is a hybrid of sorts between a peptide and a protein. It needs to be small, but it also needs to be stable, structured, and expressible by *E. coli* and yeast. This can best be described as a stabilized helix.

The McNaughton Lab has already developed a system for stabilizing helices. The basis for this system is grafting long helical sequences onto domains which stabilize extended α-helices. We have primarily used this to stabilize variants of the trastuzimab, an HIV entry inhibitor peptide based on the HIV HIV gp41 C-peptide helix. This was first done by grafting the peptide sequence onto the ~150 residue pleckstrin homology (PH) domain GLUE [181], and has been demonstrated with other PH domains [182]. More recently, the PH domain ELMO has been used to facilitate high-throughput screening via yeast display of C-helix sequences with improved binding in an HIV-entry analogue system [183]. These helical peptide mimics inhibit HIV infectivity, and are readily purified or displayed on yeast [181–183].

Our helix-stabilizing domain *du jour* is Sac7d, a hyperthermophilic chromosome binding protein [184]. Sac7d is also the basis for the anti-immunoglobulin proteins known as affitins [185], and my general strategy was to graft the helical sequence of E6AP onto Sac7d.

A.2.2 Screening System Design

Yeast Display of Sac7d-E6AP

I planned to use yeast display as my screening system due to both my familiarity (Chapters 2 and 5) and our lab's prior success in using it to perform affinity maturation of helical sequences [183]. As such, I planned to use the Sac7d-E6AP with the *myc* tag on the C-terminal. This was a difficult decision, but my rationale was that the C-terminal of E6AP was already disordered, so a linker and a small tag shouldn't interfere with folding. Confirmation of display would be easily established using the C-terminal *myc* tag (as in Section 2.3.1) with more certainty than an N-terminal *myc* tag. Successful display would then serve as a good indication that the fusion protein would be amenable to expression and purification in *E. coli*.

The DNA and protein sequence of Sac7d-E6AP for yeast display can be found in Section C.5.2, and a graphical overview of the yeast display system, with the various fluorescent readouts can be found in Figure A.3.

Yeast Display of E6AP Peptide

I also wondered if the Aga2 protein itself might not be a credible helix stabilizer, and as such, decided to also create a fusion of E6AP without any connection to the Aga2 protein on its N-terminal and the *myc* tag on its C-terminal other than the short linkers generally involved in yeast display.

Such a system would have the disadvantage of being less helically stable—even if the wild-type E6AP is a stable helix, the possibility exists that a mutation which is beneficial for binding would destabilize the helix into disorder, while a stabilized helix may be more tolerant to mutations that are beneficial for binding, but detrimental to helix formation. The advantage of a non-fusion E6AP would be the lack of the steric bulk of the Sac7d protein, as well as toleration

of the possibility that non-helical folds may be important for the docking and binding process, even if the end result of this binding process is a helix.

A.2.3 Purification of sfGFP-E6

A secondary challenge which needed to be addressed was the second half of the interaction—the E6 protein. Though there was no plan to use this as a basis for a library, and though it is a reasonably large (~150 residues) structured protein, it does not purify well due to extensive aggregation [173], and is not detectable via flow cytometry.

The aggregation problem had been at least partially solved by the substitution of four cysteine residues (C80, C97, C111, C140) for serine. The resulting protein is known as E6 4C/4S (any mention of E6 henceforth can be assumed to be this variant), and is less prone to aggregation [186], but still functional in *in vitro* assays [187], and was used to determine a crystal structure of the complex [179]. My concern was that this variant of the protein was generally expressed as a fusion with glutaione-s-transferase [186], which would add a great deal of bulk, and would not solve the issue of invisibility via flow cytometry.

My proposed solution to *both* of these problems was to fuse E6 to super-folder Green Fluorescent Protein (sfGFP), a highly stable variant of GFP [188]. This domain would serve to both stabilize E6 and as a means to detect it via flow cytometry.

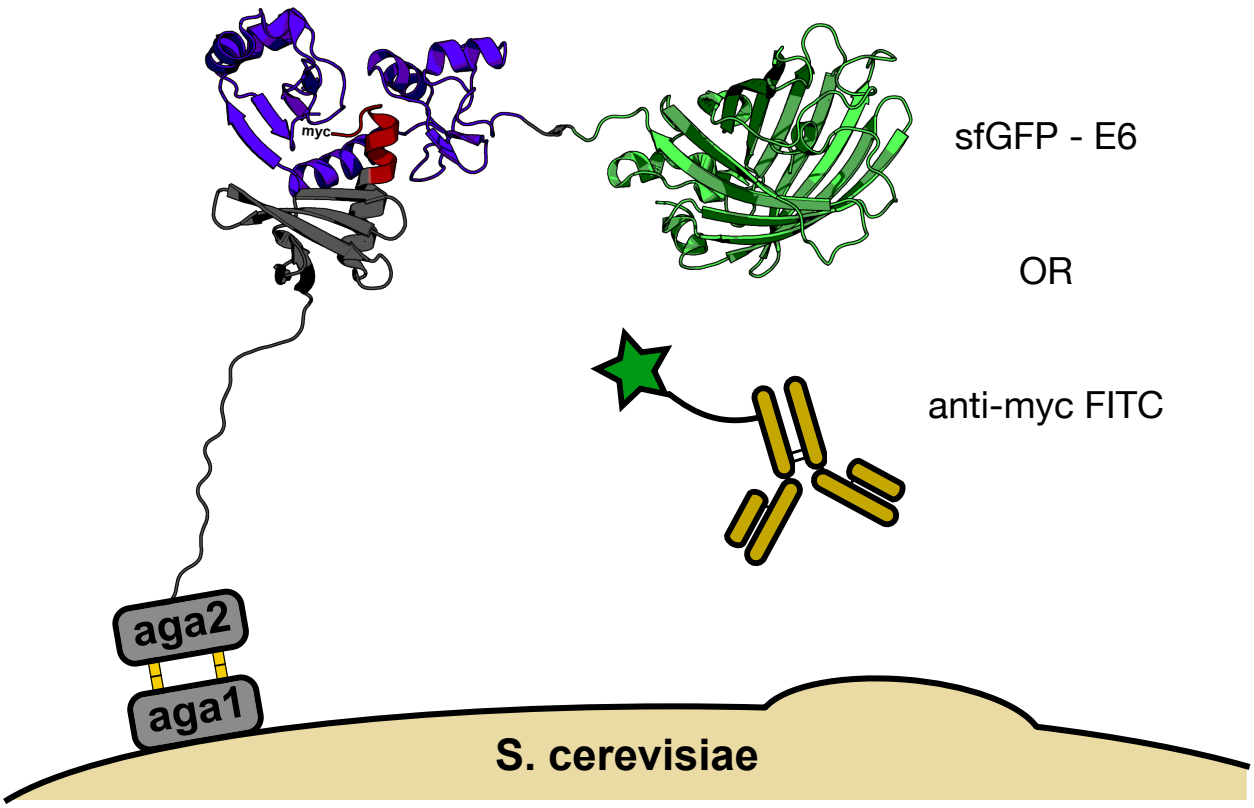


Figure A.3: Proposed Yeast Display System for Measuring Binding of E6 to E6AP The proposed yeast display system to measure binding of sfGFP-E6 to Sac7d-E6AP is shown (though it would also be used with yeast displaying E6AP peptide not fused to Sac7d). Fluorescent readouts came from either FITC (conjugated to anti-*myc* antibody) or sfGFP (fused to E6).

A.3 Grafting Strategy

The grafting strategy was developed to maximize helical stabilization and minimize steric hindrance. Since the E6AP is bound to E6 on all sides, I could not graft the helical face of E6AP onto the portion of the Sac7d helix which is directly stabilized by the β-face. I also wanted to ensure that these β-sheets would not cause a steric clash with E6 upon binding of E6AP. I used the "align" command in PyMol 2.0 (Schrödinger) on the backbones of the helical portion of Sac7d and E6AP using in order to decide where to graft the E6AP helix onto the Sac7d helix. An illustration of the final graft can be seen in Figure A.4.

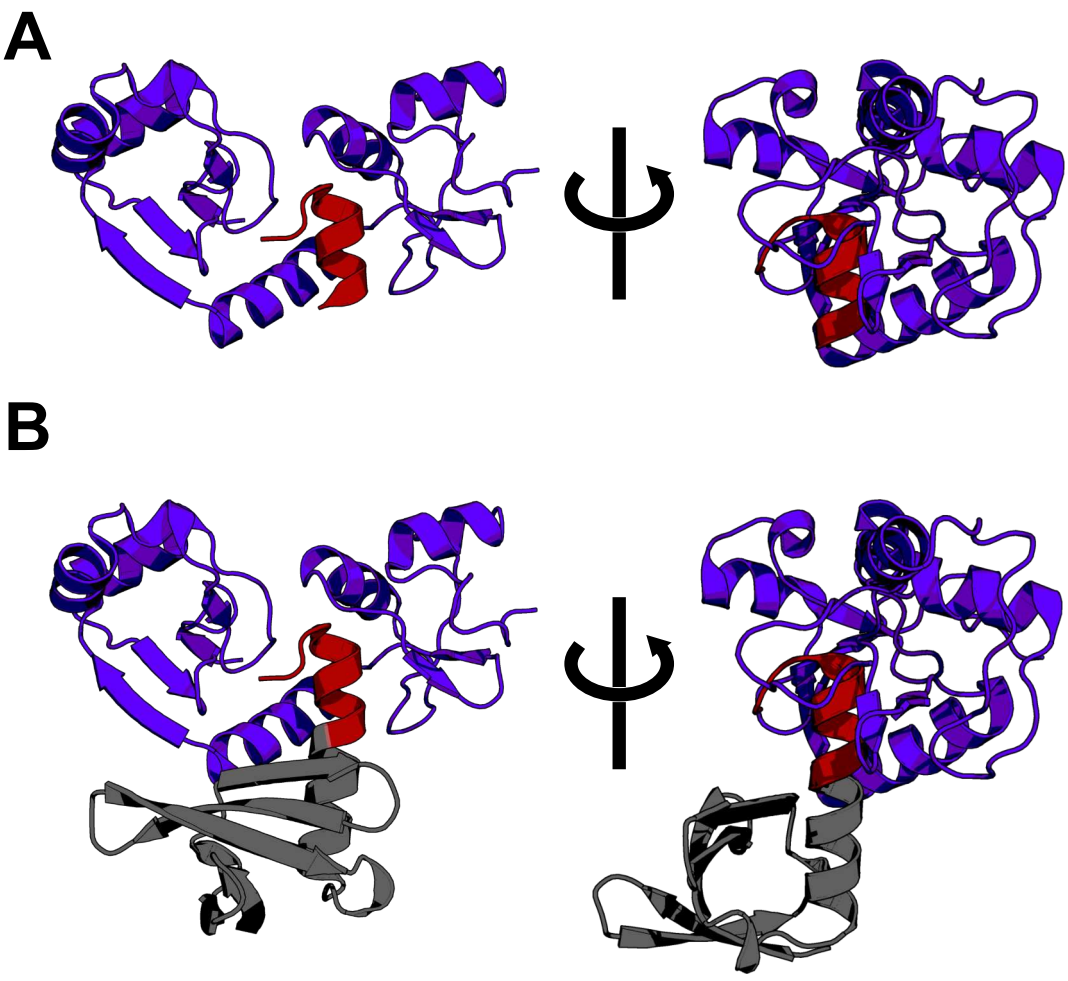


Figure A.4: E6/E6AP complex vs. E6/Sac7d-E6AP Complex Views of **A** the crystal structure of the E6/E6AP complex and **B** the same with the E6AP helix grafted to Sac7d. E6 in purple, E6AP in red, Sac7d in gray. Care was taken to minimize possible steric clashes between E6 and Sac7d.

My final decision was to graft the E6AP helix starting at position 61 on Sac7d (the residue at position 60 is from wtSac7d, while the residue at position 61 is from E6AP, sequence in Section C.5.2).

A.4 Materials Preparation and General Methods

A.4.1 Cloning

Cloning was performed as described in 2.4.1.. Any proteins for expression were ligated into a pETduet vector cut from BamHI to KpnI (thereby making the vector a pET vector with a single T7 promoter and terminator). Using the BamHI cloning site enabled use of the N-terminal His_6 tag on pETduet, which seemed desirable given the importance of the C-terminal E6AP sequence being studied. All sequences available in Section C.5.

Sac7d and Sac7d-E6AP

All Sac7d-containing proteins in this work had a standard L33T mutation, which turns off the antibody-binding property of the protein used in affitins [189]. Sac7d fusions were prepared by PCR, with Sac7d overlap, the entire E6AP sequence (N-ELTLQELLGEEER-C), and the appropriate stop codons and cloning sites.

E6AP for Yeast Display

The construct containing *E6AP Peptide-myc* (E6AP for Yeast Display) was prepared by direct insertion. Two overlapping oligos, E6AP FWD (5′-CTAGCGAGCTGACTAAACAAGAACTTC TGGGCGAGGAGCGCG-3′) and E6AP Rev (5′-GATCCGCGCTCCTCGCCCAGAAGTTCTTG TTTAGTCAGCTCG-3′) were ligated into pCTcon2 digested with NheI and BamHI (but not CIP treated), using a 50:1 ratio of insert to vector.

sfGFP-E6

The insert for the construct containing *sfGFP-E6* was created via overlap PCR. Initial PCR created an sfGFP amplicon with E6 overlap at the 3′ end, as well as an E6 amplicon with sfGFP

overlap at the 5′ end. The two amplicons were given ten cycles of PCR without primers, followed by addition of FP from the sfGFP amplification and the RP from the E6 amplification. The resulting amplicon was digested with NcoI and KpnI and ligated into pETduet.

p53 Core

The construct for *p53 Core* was generated by PCR using an IDT gBlock as a template. I used the same BamHI and KpnI cut sites as I did for the rest of the constructs in this chapter. The sequence I purified is canonically positions 94–292 of p53. Though the protein was never used in an assay, it purified well from BL21 *E. coli* with normal induction (0.5 mM IPTG, 16 hours, 25 °C). The DNA and protein sequences can be found in Section C.5.8.

A.4.2 Protein Purification

Proteins were purified using a standard nickel/His_6 protocol, as described in Section 3.2.3, and quantified by absorbance at 280 nM. Sac7d, Sac7d-E6AP, and sfGFP purified without issue and remained stable for at least two weeks following purification. The purification of sfGFP-E6 proved more problematic, as can be seen in Figure A.5.

Finalized conditions for this the induction are: initiation of induction at OD_{600} = 0.6 with 0.5 mM IPTG, induced for 16 hours at 30 °C with 1 L of LB in a 2 L erlenmeyer flask, shaking at 250 rpm. As can be seen in Figure A.6, the protein purifies at usable purity under these conditions, and remains stable for a usable amount of time.

A.4.3 Protein Purification Assay Consequences

One note that pertains especially to ITC (Section A.6) is that Sac7d and Sac7d-E6AP would typically come out of dialysis at a concentration of ~10–100 μM, and would require concentration after dialysis and before ITC. This is probably inconsequential for the yeast display assays (Section A.5,Figure A.9), but it is a problem for ITC. Though this *shouldn't* lead to a buffer mismatch, it often seemed to do so. Dialysis at high concentration, however, generally led to significant material loss.

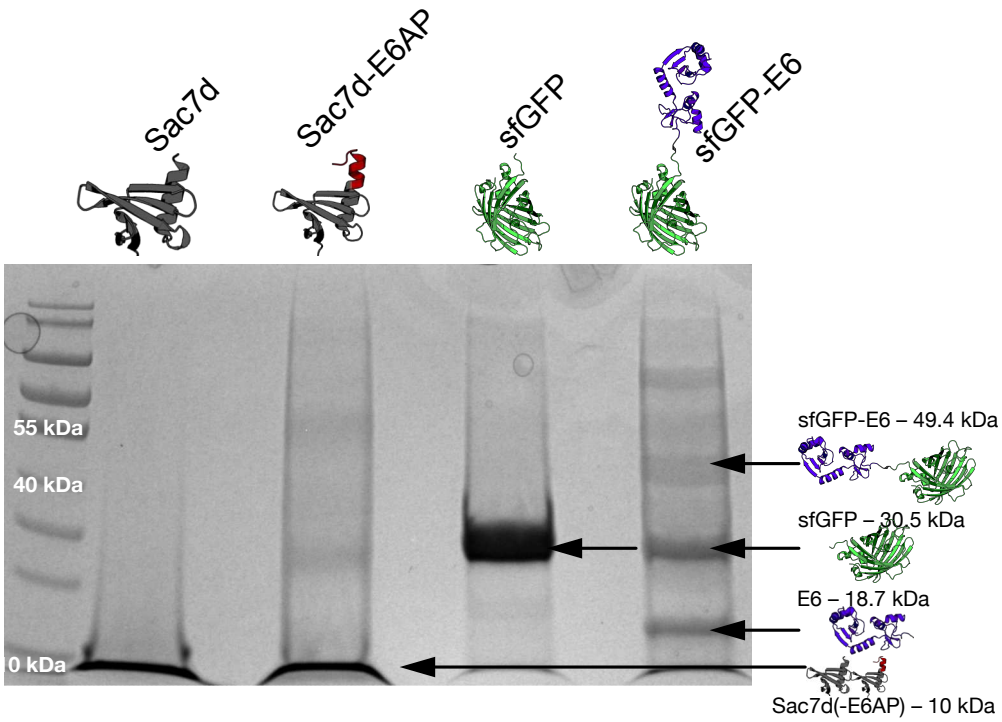


Figure A.5: Initial Purifications of Sac7d(-E6AP) and sfGPF(-E6) This protein gel, run about a week after protein purification as a quantitation control demonstrates the tendency of the sfGFP-E6 protein to degrade over a short period of time.

A.4.4 Yeast Preparation

Yeast were transformed via electroporation (as in 2.4.1), and induced via normal galactose induction (as in Section 2.4.2). Inductions were kept short, never longer than 24 hours, and were performed at room temperature (~25 °C). Induced yeast were kept at 4 °C for 1–3 days prior to flow cytometry.

Flow cytometry was performed on a CyAn ADP, unless otherwise specified. Following incubation conditions listed with the relevant data, cells were washed once and left as pellets on ice. Flow cytometry was performed by re-suspending the pellets in 1 mL of PBS-BSA. Front and side scatter were used to gate healthy cells (typically >90% of the total), and the fluorescence of 50,000 healthy yeast cells was measured.

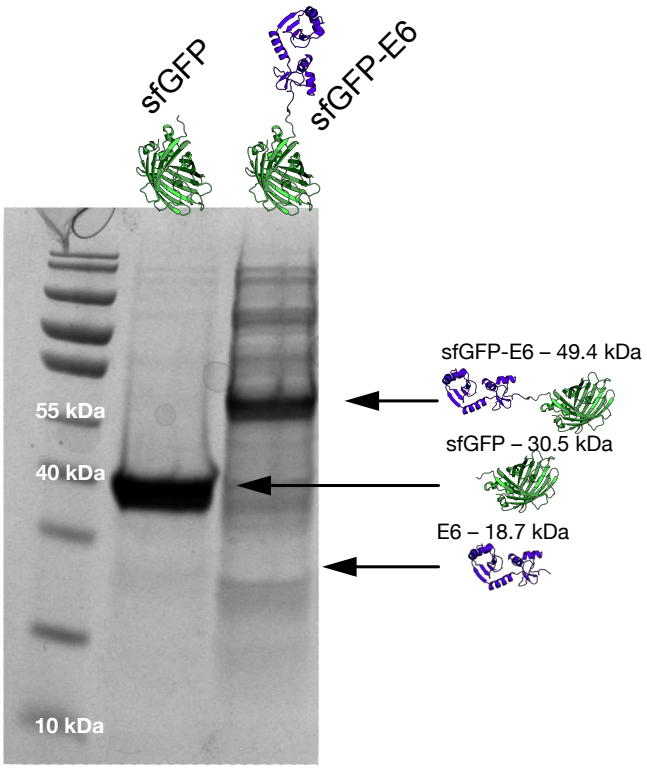


Figure A.6: Final sfGFP-E6 Purification PAGE gel run ~1 day after purification demonstrates that under the right conditions the sfGFP-E6 fusion purifies reasonably cleanly, and is stable over short time periods.

A.5 Yeast Display Assays

A.5.1 Sac7d-E6AP and E6AP Display on Yeast

In order to analyze the efficacy of our system for measuring binding between displayed Sac7d-E6AP or E6AP and E6, I first had to ensure that they displayed on yeast, and that the sfGFP fluorophore proposed to track E6 binding did not generate a false-positive signal.

I incubated ~10^6 induced yeast in PBS-BSA with a 1:1000 dilution of FITC conjugated anti-*myc* antibody or (abcam ab117599) 10 μM sfGFP for 45 minutes rotating at 4 °C. Cells were washed once and measured for FITC/GFP fluorescence with a CyAn ADP flow cytometer. Figure A.7 shows these results.

A.5.2 Initial Tests of sfGFP-E6 binding to Displayed E6AP

After establishing that the proteins of interest would display, and had only minimal affinity for sfGFP (Figure A.7), I decided to determine whether the Sac7d-E6AP and E6AP peptide would bind sfGFP-E6. Yeast cells were again induced for 24 hours, and I initially set up the experiment in a similar fashion to a normal yeast display experiment—500 μL of PBS-BSA for 45 minutes at 4 °C with 5 μM protein (as well as the usual display checks with FITC conjugated antibody and negative controls). The results of this experiment are shown in Figure A.8.

It was obvious that *something* was occuring, and I had both reason to be hopeful and skeptical. The hopeful signs were that neither the yeast displaying Sac7d (without any E6AP fusion) seemed to have any appreciable affinity for the sfGFP-E6, while both means of displaying E6AP did. However, the yeast displaying Aga2 alone had comparable signal to the two E6AP displaying samples. Given future results, my hypothesis is that the Aga2 control yeast are displaying more copies of the Aga1/Aga2 pair, and that these Aga pairs are more sterically accessible. Given that the pair is held together by disulfide bonds (see Figures 2.2 and A.3), it is possible that the cysteine-rich E6 protein has some affinity for the Aga1/Aga2 proteins (or possibly mis-folded copies), which are held together by disulfide bonds. This effect would intuitively be especially pronounced when they are numerous and sterically accessible.

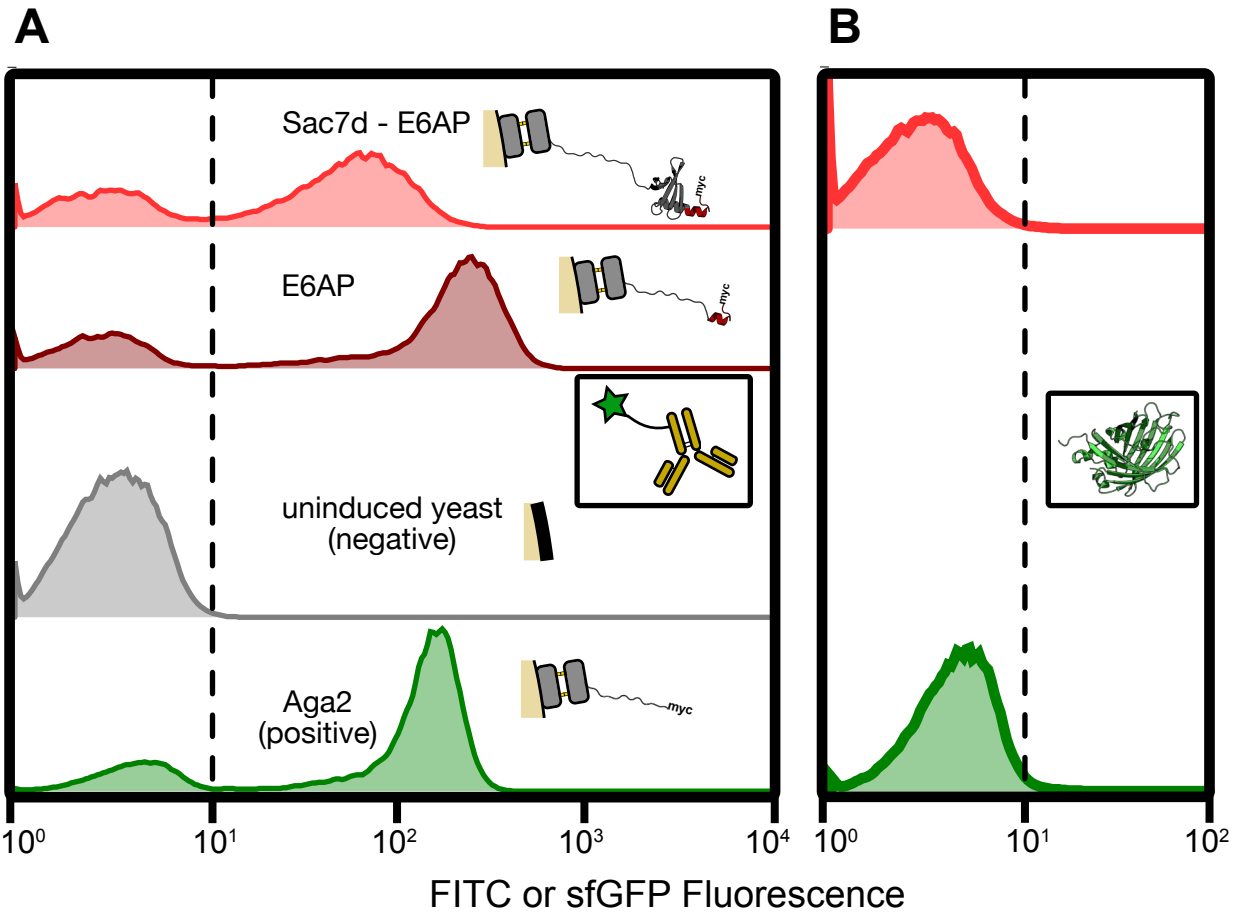


Figure A.7: Initial Confirmation of Sac7d-E6AP and E6AP Display on Yeast The initial experiment checking for display of Sac7d-E6AP and E6AP **A** demonstrated that both the fusion and the isolated peptide display well on yeast, with 52.4% of the yeast in the Sac7d-E6AP sample and 58.8 % of the E6AP peptide displaying their proteins, which compared well to the positive Aga2 control (74.3% positive) (cutoff indicated by dotted line). Under the same conditions, an uninduced sample of yeast bearing the *Sac7d-E6AP* plasmid were <1% positive. Additionally, I needed to ensure that sfGFP did not generate a positive signal even absent the E6. As can be seen in **B**, it does not, with only 3.8% of Sac7d-E6AP and 2.8% of Aga2 displaying cells generating a positive signal.

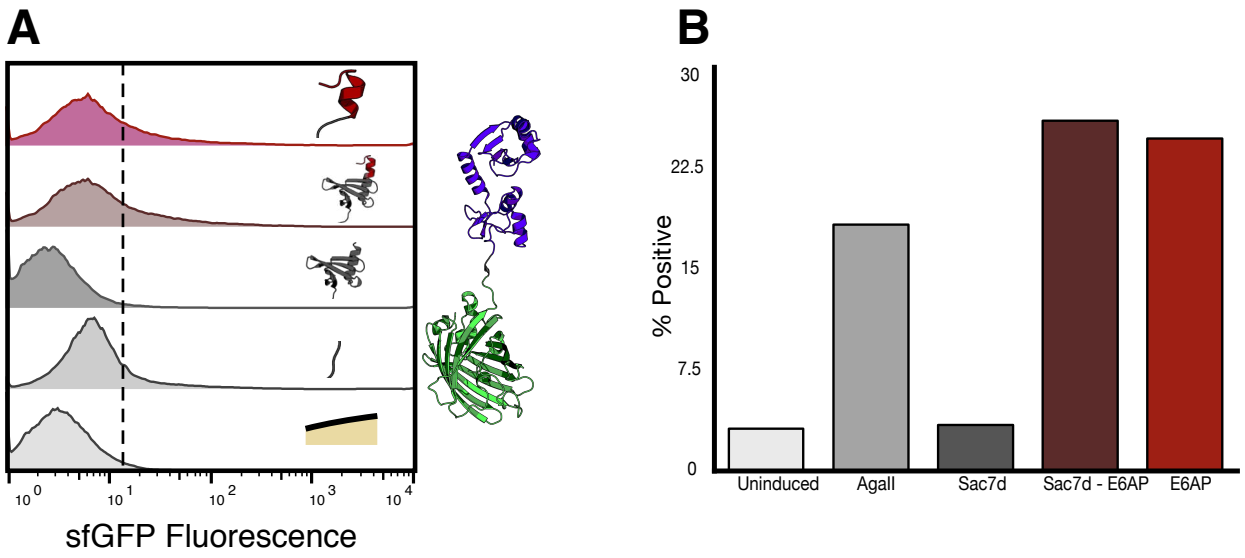


Figure A.8: E6 Binding by Displayed E6AP, 45 min. Incubation Initial experiment **A** histograms of sfGFP fluorescence showed no obvious positive population, but did display a "shoulder" from the negative populations. The bar graph in **B** shows the % positives resulting from these populations.

A.5.3 sfGFP-E6/E6AP Binding with Longer Incubation Times

Given that I suspected that E6/E6AP binding was occurring (and the results were not simply the result of artifacts), and that these binding events took the form of a "shoulder," I wondered if binding was *beginning* to occur, but had not yet achieved an equilibrium after <1 hour.

To test this, I tried a longer incubation. I still used normal incubation conditions—10^6 cells in 300 µL of PBS-BSA, with either 10 µM sfGFP-E6, 30 µM sfGFP, or 1:1000 FITC-conjugated anti-*myc* antibody—but I incubated them for 20 hours at 4 °C. Negatives and binding controls were normal, and the sfGFP-E6 binding data can be seen in Figure A.9.

Though these data are not conclusive, they indicate that given sufficient time, E6AP binding to E6 will occur. Though I did not have the opportunity to properly replicate this experiment, and so cannot be sure that this is not due to some artifact or fluke of induction, it seems unlikely that the two E6AP samples would have the same fluke, while the others do not.

Given that binding is good, but not great (2.3 ± µM as measured by SPR in [190]), and the accepted measurements use relatively high concentrations, it is possible that our more complicated yeast display setup simply requires more time to equilibrate. Since our system is more sterically

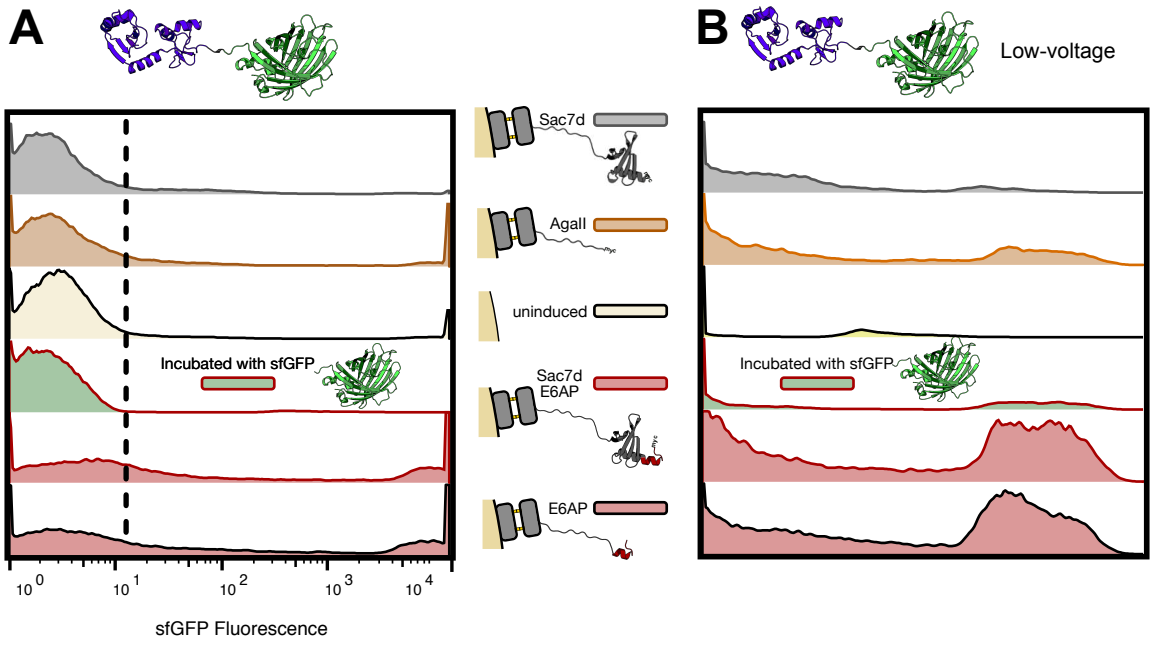


Figure A.9: E6 Binding by Displayed E6AP, 20 hr. Incubation After 20 hours of incubation with sfGFP or sfGFP-E6 there was a clear positive population. **A** The difficulty was that this population was off the scale, so I re-adjusted the voltage (thereby making the instrument less sensitive) so that all events were on scale, and **B** ran the samples again. Both histograms also contain one Sac7d-E6AP sample incubated with 30 µM Though quantitating this improvement is difficult due to the continuing presence of a "shouler" from the negative samples, but ~30% of E6AP/Sac7d-E6AP samples are positive compared to the ~10% of the comparable Aga2 control.

hindered, and uses more structured peptides (thus, possibly structured in a sub-optimal configuration), it is possible that the interaction takes longer to occur, but is still robust. I also suspect that there is some issue with the folding of the sfGFP-E6 fusion when it is not bound to E6AP, meaning that our effective reagent concentration is lower than expected.

In any case, though not conclusive, these data demonstrate the high likelihood that yeast display is a good platform on which to simulate peptides based on E6AP (both separately and as a Sac7d fusion), and it may be possible to use this platform to select for variants with improved binding.

A.6 sfGFP-E6/E6AP Binding via ITC

A.6.1 Protocol

In addition to yeast display, I also attempted to measure the E6/E6AP interaction by ITC. I opted to use the same sfGFP-E6 fusion protein as my E6 source, and purified Sac7d-E6AP (Sequence: C.5.6). Experimental conditions are given in the appropriate figures, but ITC experiments were performed on an ITC200, with 10× concentration/2× molar quantity in the syringe. Total cell volume was 350 µL, and total syringe volume 40 µL. Experiments were performed at 25 °C with a stirring speed of 750 RPM. Reference power was 3 µcal/second. The first injection of 0.4 µL occurred after 60 seconds, followed by 15 injections of 2.49 µL at 180 second intervals.

I opted to put the Sac7d-E6AP in the syringe, since it seems to be more stable, and is certainly easier to purify at high concentrations. The data shown in Figure A.11 indicate that it may be better to reverse this, however.

A.6.2 Results

Like the yeast display data, the ITC data indicates that Sac7d-E6AP binds to E6 (in the form of sfGFP-E6), but falls short of being conclusive. The best representative data can be seen in Figure A.10.

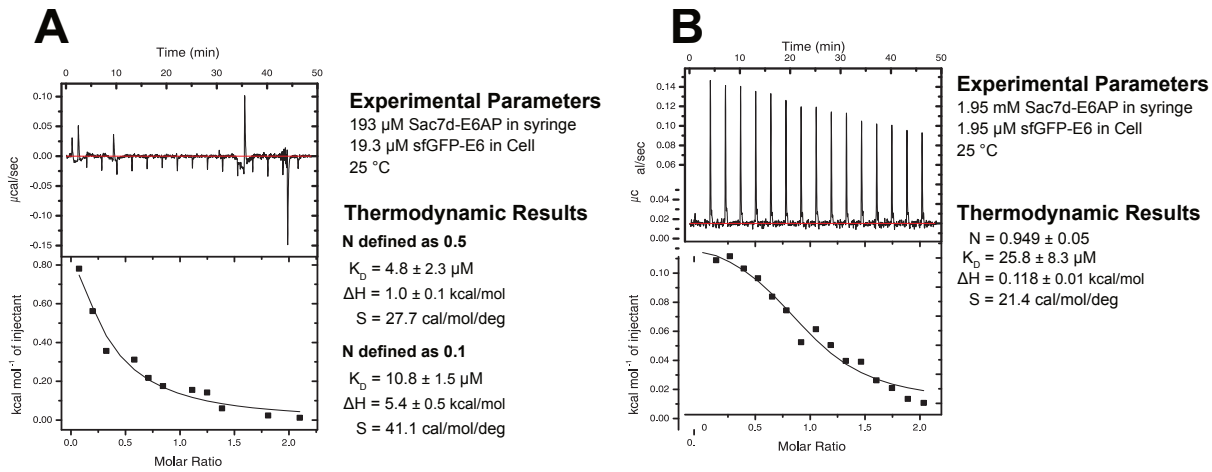


Figure A.10: ITC Titrations of Sac7d-E6AP into sfGFP-E6 In **A** an ITC titration and line of best fit is shown for a 193 µM Sac7d-E6AP into 19.3 µM sfGFP-E6, along with the thermodynamic results of two seperate fit equations. Drastic spikes can be seen in the raw ITC heat. I assume they correlate to aggregation events, and those data have been discarded. **B** shows the results of a similar titration with ~10x concentration in both cell and syringe.

I used Origin 7.0 to analyze the data shown in Figure A.10, and the fitting algorithm continually drove the N-value toward zero as best-fit iterations were performed. In order to get convergence I instructed the program to hold this value constant. This is obviously not ideal, and indicates some discrepancy in concentration measurement (either due to the protein measurement being flawed, or some protein being inactive), as well as a sub-optimal c-value. The latter, especially, is not surprising, given that (assuming the thermodynamic values are accurate) my c-value for this titration is fairly low (~2–5). The fact that varying the N-value has minimal effect on the K_D indicates that these K_D values are *reasonable*, if not precisely trustworthy.

This becomes an even more reasonable assumption given that the titration shown in Figure A.10, which is done at a more reasonable c-value of ~10, has a similar K_D . Therefore, I believe it is accurate to conclude that there *is* a binding interaction between Sac7d-E6AP and sfGFP-E6. Also worth noting is that this binding interaction seems to be *entropically* driven, and at high concentrations even endothermic.

The most obvious caveat to this is shown in Figure A.11, which puts side by-side the titration from Figure A.10B and a titration of 1.44 mM Sac7d into buffer. In this case, the Sac7d into buffer generates ~4 times *more* heat than Sac7d-E6AP into sfGFP-E6. My own supposition is that this is

due to some kind of non-covalent oligomerization in the extra-concentrated Sac7d. The fact that the heats are so drastically different between the two samples, as well as the fact that this is *not* an effect I have seen every time I've titrated Sac7d into buffer, suggest that whatever is occurring in Figure A.10B is distinct from the dilution-driven heats in Figure A.11B.

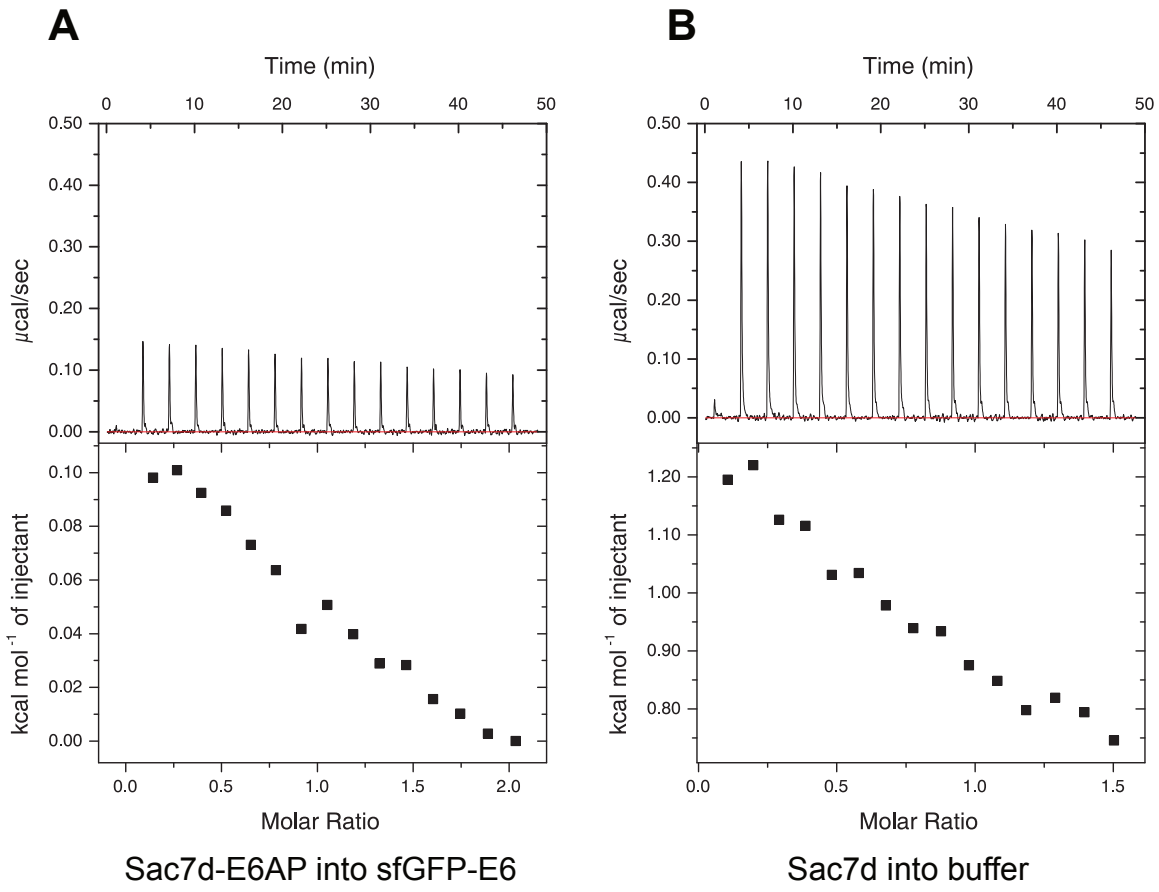


Figure A.11: ITC of Sac7d into Buffer Compared to Sac7d-E6AP into sfGFP-E6 A shows the same data seen in Figure A.10B (1.93 mM Sac7d-E6AP into 193 µM sfGFP-E6) side-by-side with **B** 1.44 mM Sac7d into buffer, with molar heats calculated as if it were being titrated into 193 µM sfGFP-E6, as a means of normalization.

Again, the corpus of data *suggest* that Sac7d-E6AP binds to sfGFP-E6, and that this binding is measurable by ITC. That said, a technique more suited to µM K_D values, such as fluorescence polarization, might be a better choice.

A.7 "Helical Grafting of E6AP" Conclusions

In conclusion, I *appear* to have succeeded at my goal of creating a stable platform for expression and study of the E6AP peptide. I hope that, if this project is advanced in the future, this platform will be used to generate better binders to E6. As it is, there is interesting functional possibility in simply replacing binding between E6 and the full E6AP protein (which is, after all, a ubiquitin ligase) with a binding interaction between E6 and the harmless E6AP peptide.

A.8 Future Directions

A.8.1 Replication

The first and most obvious future direction of this project is replication. The data in this chapter are suggestive, but not beyond reproach. The key to this, is, I believe, making sure that all materials are as fresh as possible. This is challenging, because the E6 especially is prone to degradation and aggregation, and does not purify efficiently. Related to this replication would be the determination of a K_D curve via yeast display (as in Section 2.8.3, which should confirm past ITC experiments, and guide future possibilities.

A.8.2 Yeast Display

The next step for the yeast display/flow cytometry route is an incubation time series. It has been established that ~1 hour is too short a time for a full incubation, but ~1 day is probably too long, and is likely introducing artifacts. I predict that 4–6 hours of incubation will result in differentiation between samples without the massive "off-scale" signal of the ~1 day incubation.

A competition assay utilizing multiple concentrations of non-fluorescent E6 fusion protein would help rule out the possibility of artifacts. E.g. a population of yeast displaying E6AP should have lower fluorescent signals when incubated with 5 μM sfGFP-E6 and 45 μM MBP-E6 (for instance) than they would if they were incubated with only the 5 μM sfGFP-E6. If this is *not* the case, it could be an indication that the yeast are somehow taking up the protein, rather than merely binding it on the surface.

A.8.3 ITC

The next step for ITC is to either do a more extended titration utilizing multiple refills of the syringe, or to use more concentrated materials. I believe that more concentration of materials will involve co-dialysis of the proteins together at high concentration. A dialysis at such high concentration will result material loss. Of course, such a dialysis would need to be done in series and/or at hight volume to prevent the contamination of the buffer at large with significant amounts of protein.

Alternate E6 Forms

Since purification of sufficient quantities of the sfGFP-E6 has proven to be a bottleneck, it may be worthwhile to try out other stabilizing domains (such as MBP) for ITC assays.

A.8.4 Other Assays

As a general statement, more types of assays means more confirmation, and any positive result is more believable if it is corroborated in multiple systems. I think the best choices of alternative assays for this system would be:

- Fluorescence Polarization: because it seems to work well in the μM range of this interaction, and because it can be done with constructs on hand.
- SPR: because it has already been used extensively for the E6/E6AP system [190], and because SPR will allow a *kinetic* comparison between free and stabilized peptide that should illuminate the other assays (especially incubation time of the yeast display assay)
- ELISA: because it is sensitive, robust and well-suited to binary "yes/no" determinations of binding. An anti-GFP antibody would work

A.8.5 Introducing p53 to E6/E6AP Interaction

The next *experimental* step (as opposed to technical step) would be to probe the system in the presence of p53. The assays would be similar to those already performed, but with co- and

pre-incubation of p53. Since flow cytometry seems to be the most promising avenue at the moment, obtaining the p53 with an orthogonal fluorophore, and/or using Förster resonance energy t (FRET) pairs would enable determination of *correlation* between E6 and p53 binding events.

Appendix B

Other Experiments of Possible Interest

Very few theses are exhaustive lists of a graduate student's time, the main body of this one probably doesn't even manage a simple majority of my time and energy. As such, I wish to devote a few pages to some of the side projects related to the work in Chapters 1-5 (Sections B.1-B.3), as well as a few independent projects .

B.1 Binding of TBP 6.7 to ΔC25 TAR RNA

B.1.1 Introduction

The affinity of TBP 6.7 for a canonical TAR sequence (with a trinucleotide bulge) has been well-established (Chapter 3). But the HIV-1 TAR Element does have some variation in the wild [39], and the canonical TAR element in HIV-2 is a dinucleotide bulge [191].

As part of a grant application, I tested the affinity of TBP 6.7 for a variant of TAR with a *di*-nucleotide bulge, designated as TAR ΔC25. There was never a good publication in which to include it. It was temporally separate from the data in Figure 3.9, and has additional concentrations, that I decided to not simply blend the data.

B.1.2 Methods

The methods used are essentially identical to those used to gather the data in Figure 3.9, and are described in detail in Section 3.3.1.

The TBP 6.7 sequence can be found in Section C.2.3. I once again used the C-terminal His_6 and FLAG Tag variant.

The TAR ΔC25 (5′-biotin-GCAGAUCGAGCCUGGGAGCUCUCUGC-3′ also lacks the extensive G·C clamp at the 5′/3′ extremes, but is otherwise identical to the canonical TAR used previously. An illustration is found in Figure B.1.

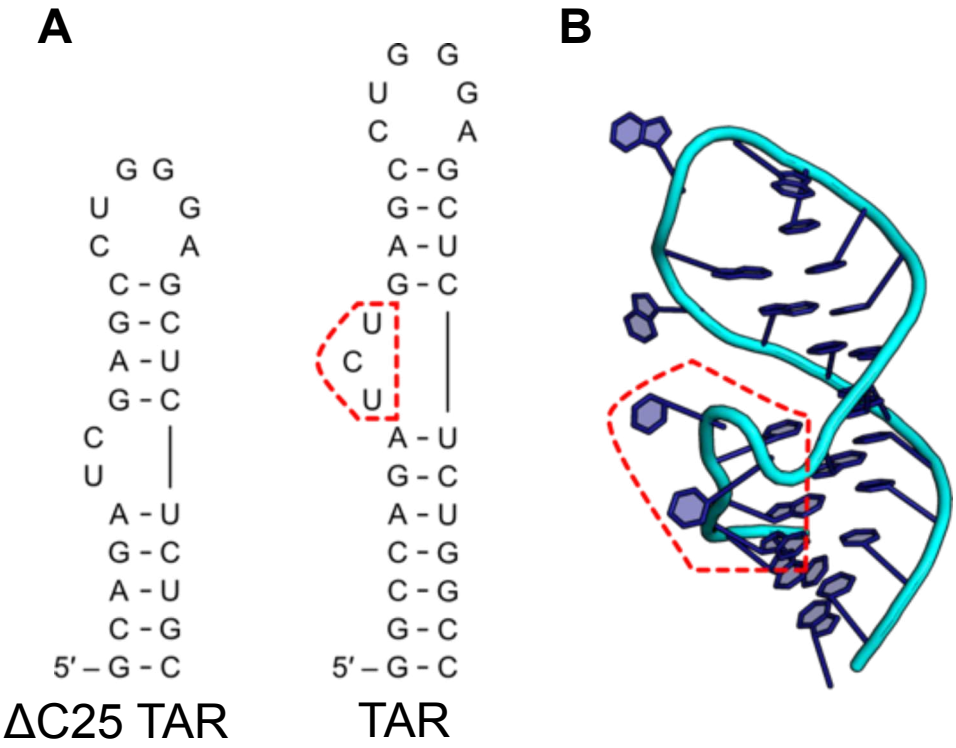


Figure B.1: Dinucleotide Bulge TAR The sequence and putative secondary structure **A** of the ΔC25 TAR analyzed here as well as the sequence of TAR used extensively in Chapters 2-5. Shown in **B** is a crystal structure of TAR with the trinucleotide bulge outlined in red, demonstrating its significance to the overall structure.

In the experimental design TAR was used as a positive control, and a bulgeless variant (hp1 from Figure 3.9) was used as a negative control.

B.1.3 Results

TBP 6.7 binds ΔC25 TAR nearly as well as it binds the canonical TAR used throughout this thesis. At a protein concentration of 2 nM, the ELISA signal for ΔC25 TAR was somewhat less than half of the signal for TAR (which is still quite high). At 20 nM, the signal was ~75%, and at 100 nM was essentially identical. Based on these data, I would estimate a K_D of around 10 nM, ~1 order of magnitude worse than the affinity for TAR, but still quite good.

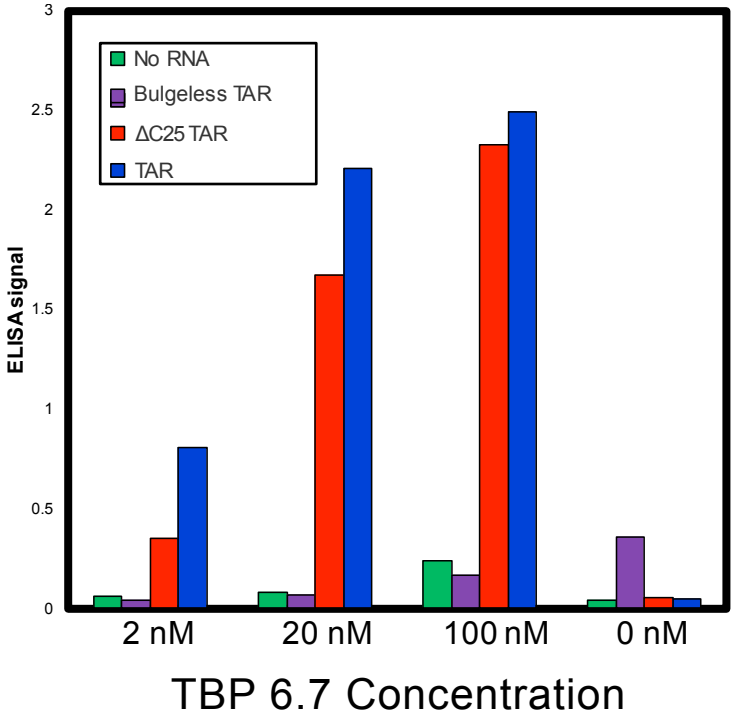


Figure B.2: Binding of TBP 6.7 to ΔC25 TAR TBP 6.7 binds ΔC25 TAR nearly as well as it binds the TAR sequence used throughout this thesis.

B.1.4 Conclusions

This outcome was unexpected when considering the crude mutation data shown Figure 3.9 because TBP 6.7 binding is so sensitive to the presence of the trinucleotide bulge. However, it

is less surprising in the face of the crystallography data, which indicate that TBP 6.7 does not engage much with the bulge directly, but rather that the bulge helps form the major groove that TBP 6.7 binds in a double-stranded mode (Chapter 4). A dinucleotide bulge has a similar effect in this sense as a trinucleotide bulge does, and therefore similar affinities.

What is most promising is that TBP 6.7 engages with TAR under approximately the same conditions as Tat itself does. That is to say that mutations which would result in abolishment of TBP 6.7 binding would also likely result in a lack of Tat/TAR-dependent transcription activity. For instance, the Tat peptide from HIV-1 (which recognizes a trinucleotide bulge-containing TAR element) is able to transactivate the dinucleotide bulge (similar to ΔC25 TAR) from HIV-2

B.2 Alternate β2β3 Loop Display Strategies

B.2.1 Introduction

The data presented in Section 5.8.1 was only a part of an ambitious program to display variants of the TBP 6.7 β2β3 peptide on bacteria. I attempted to create a cysteine-flanked version, and synthesized a molecule—6,6′-sulfonylbis(1,2,3,4,5-pentafluorobenzene)—which had been used to create similar cyclic libraries on bacteriophage [192].

However, such a library would only be useful if it could be shown that the conjugation had worked robustly, which would be best demonstrated by the presence of free thiol groups prior to the conjugation, and the absence of same following conjugation.

Two basic methods were attempted to determine the viability of this plan: a mass spectrometry probe based on TEV cleavage (Figure B.3A), and a flow cytometry based probe based on FITC conjugation (Figure B.3B). To enable these tests, I constructed a version of the TBP 6.7 β2β3 loop (loop only, analogous to *peptide* **1s** from Chapter 5) flanked by cysteines and TEV cleavage sites, fused with eCPX for display on *E. coli*.

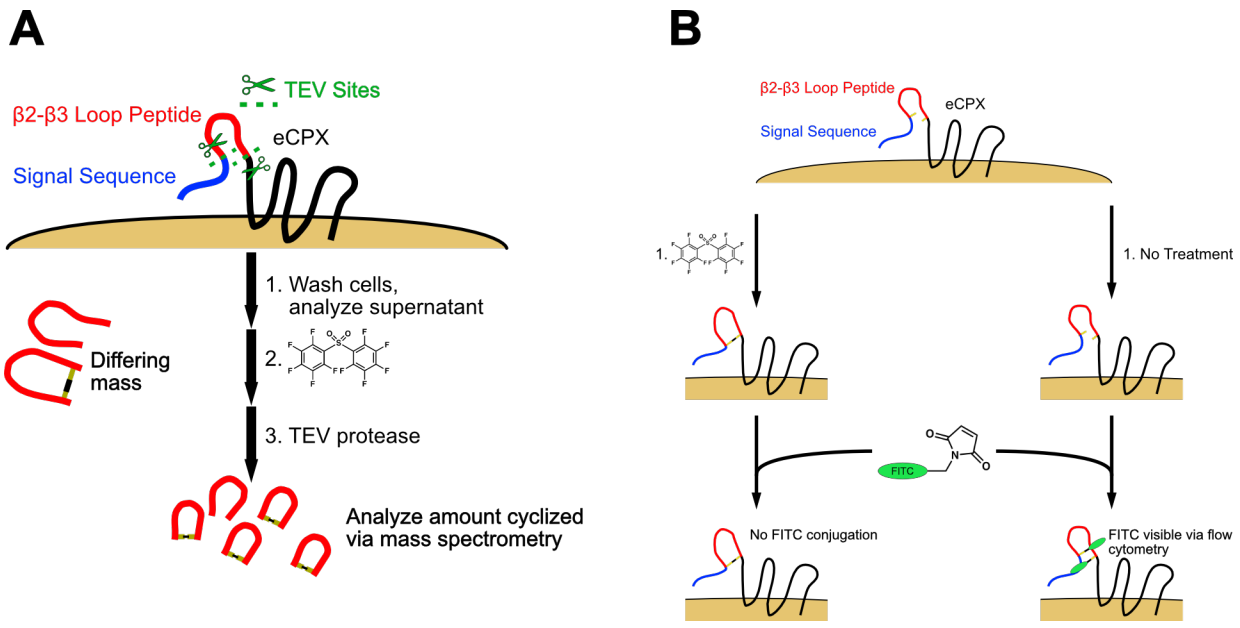


Figure B.3: Schemes for Detection of Cyclization of Displayed β2β3 Loop Peptide The proposed schemes for confirmation that the β2β3 loop library had indeed been cyclized via **A** TEV cleavage with differentiation between cyclized and uncyclized via mass spectroscopy, and **B** fluorophore conjugation detected via flow cytometry.

B.2.2 TEV-cleavage displayed β2β3 loop peptides

I tested the feasibility of using TEV cleavage to determine whether cyclization had occured. An additional benefit of this assay is that it could potentially be used in a pulldown assay, or manipulated for use in assays using cleavage-dependent logic.

Methods and Results

TEV protease was purified per standard purification protocols.

Bacteria displaying either TEV-Cys-β2β3-Cys-TEV or Z-peptide (used as a negative control, sequence from [193]) as eCPX fusions were inoculated and induced according to the protocol in Section 5.8.1. After a 1 hour incubation with FITC-conjugated anti-*myc* antibody, they were measured via flow cytometry and found to be displaying at high levels (Figure B.4)A. ~10^9 cells were then incubated with ~1 μM TEV protease in 1 mL of PBS for ~16 hours, and once again tested for display by use of FITC-conjugated anti-*myc* antibody.

Subsuquently, a similar experiment was performed with 10× the number of cells in 10× the volume. After the initial centrifugation, the supernatant (which should contain β2β3 peptide for the TEV-Cys-β2β3-Cys-TEV sample) was saved, lyophilized to reduce volume ~10-fold, and analyzed for presence of peptide via mass spectrometry.

This incubation with protease had little effect on the autofluorescence of the cells (Figure B.4B), nor the display of Z-peptide (Figure B.4C), but reduced display of TEV-Cys-β2β3-Cys-TEV by ~⅔ (Figure B.4D).

However, it was not possible to detect the cleaved β2β3 in the supernatant via mass spectrometry (Figure B.5) even with a 10-fold increase in bacteria (for a total of 8.8 × 10^9 cells in a 1 mL reaction. This is perhaps not surprising, since assuming 10^5 displayed proteins on each cell, there would only be a small amount of peptide present, even at 100% cleavage—about 15 μMafter lyophilization. See following equation.

$$\frac{8.8\times10^{9}cells}{10^{-3}L}\times\frac{10^{5}Peptides}{cell}\times\frac{1mol}{6.022\times10^{23}Proteins}\times\frac{10^{6}\mu mol}{mol}=1.5\mu MPeptide$$

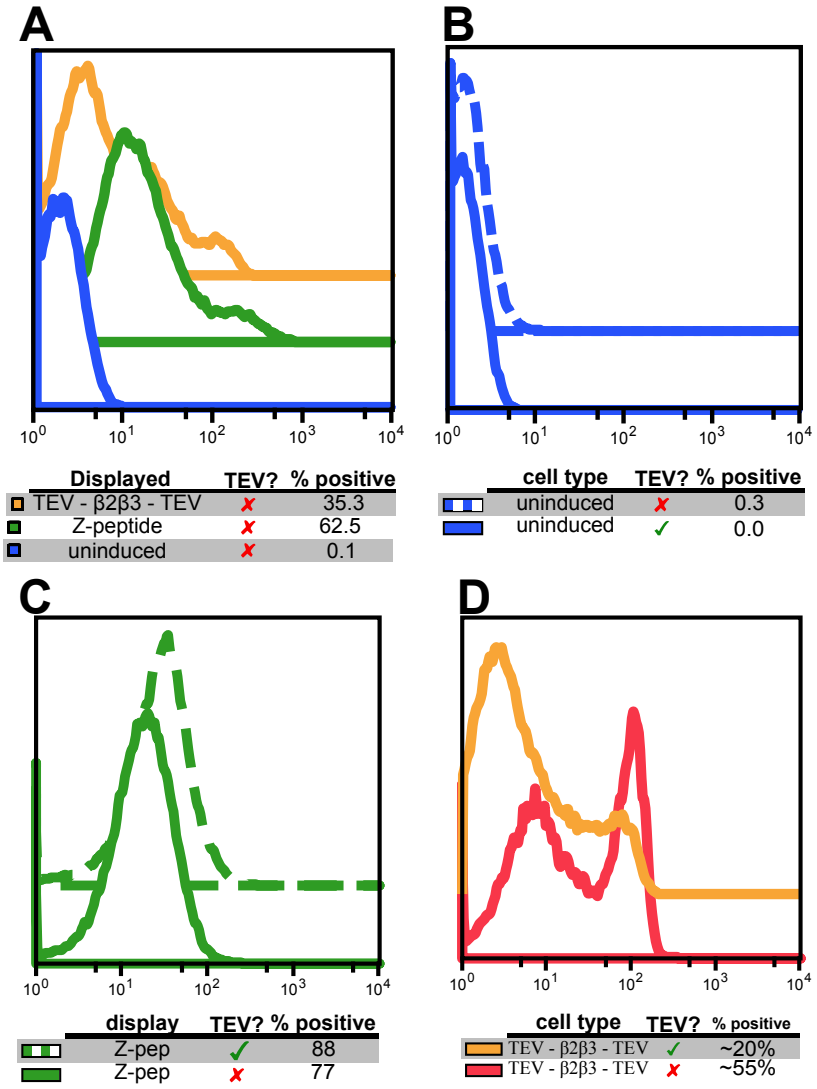
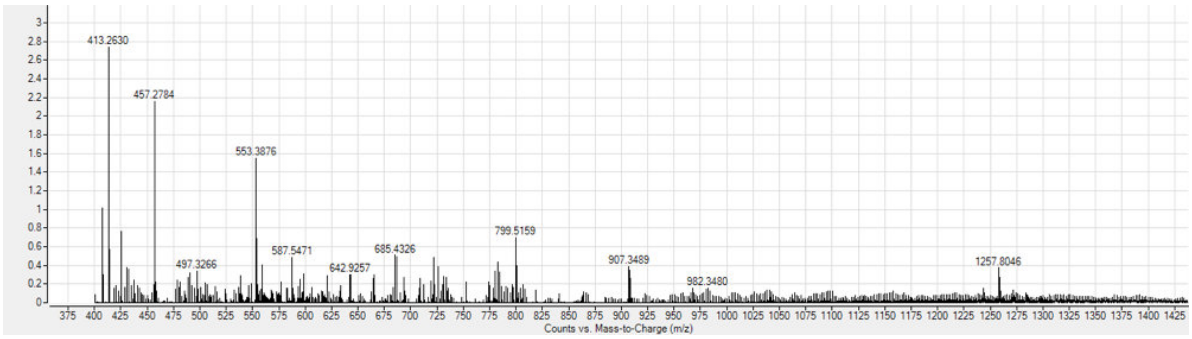


Figure B.4: Display of β2β3 Peptide Variants Before and After TEV Cleavage **A** shows the levels of display prior to TEV cleavage. Following ~16 hours of incubation with ~1 μM TEV protease **B** The uninduced cells and **C** the cells displaying Z-peptide were largely unaffeced, while **D** a great deal (~⅔) of the TEV-Cys-β2β3-Cys-TEV had apparently been cleaved away.



Expected Mass of β2β3 loop peptide = 2410 Da

Figure B.5: Mass Spectrum of Supernatant Following TEV Cleavage There is no obvious peak for the β2β3 peptide in this spectrum, with only the peaks at ~800 Da and ~1257 Da possibly corresponding to some version of a Z=2 and Z=3 β2β3 peptide post-cleavage.

B.2.3 Maleimide FITC Conjugation of TEV-Cys-β2β3-Cys-TEV

Methods and Results

I inoculated and induced bacterial cells displaying either a Z-peptide eCPX fusion (as a negative control), or Cys-β2β3-Cys. Display was confirmed via incubation with FITC conjugated anti-*myc* antibody.

I then tested the using the number of free thiols by incubating (for 1 hr. with 500 μM maleimide-FITC in PBS) the Z-peptide cells, Cys-β2β3-Cys cells that had been incubated for 30 min. with 2 mM of the reducing agent tris(2-carboxyethyl)phosphine (TCEP), and Cys-β2β3-Cys cells that had not been treated with TCEP. Results are shown in Figure B.6.

Given that the the most positive sample was that displaying Z-peptide, which does not have any cysteines (Figure B.6A), this does not appear to be a promising method for identifying the presence of free thiols on a displayed protein. I suspect this is due to the fact that the bacterial surface itself has numerous membrane proteins containing accessible cysteines. This assertion is supported by the fact that the "negative" population of the cells treated with TCEP (Figure B.6B) is clearly shifted to the right compared to the cells that were not treated with reducing agent (Figure B.6C), which would occur if the membrane proteins of cells *not* actively displaying eCPX fusion were being reduced. It is also likely that the porin-like eCPX protein facilitates diffusion of the relatively small maleimide-FITC into the cell interior.

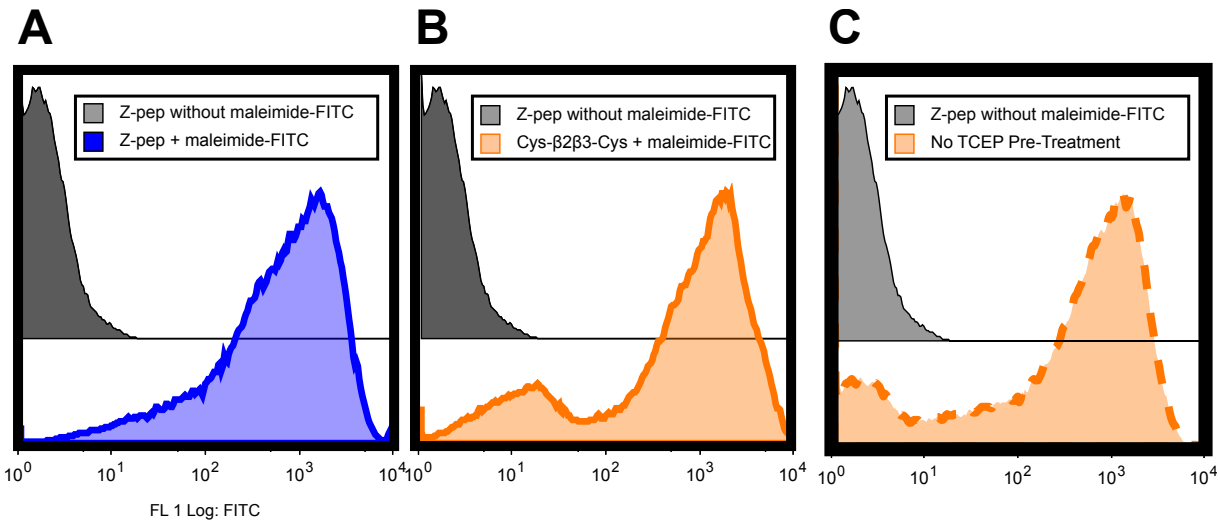


Figure B.6: Results of Incubating Cys-β2β3-Cys with Maleimide-FITC A negative control, consisting of bacteria displaying Z-Peptide, but not incubated with maleimide-FITC is shown against the results of incubating maleimide-FITC with **A** bacteria displaying Z-peptide (91.6% positive) **B** Bacteria displaying Cys-β2β3-Cys after pre-treatment with TCEP (81.6% positive) and **C** Bacteria displaying Cys-β2β3-Cys, but not treated with TCEP (76.6% positive)

B.2.4 Conclusions and Future Directions

Conclusions

Though we chose not to pursue these avenues further, there are some promising results here. Most notable is that both TEV-Cys-β2β3-Cys-TEV and Cys-β2β3-Cys do, in fact, display on bacteria, even with their cysteines.

The results of the assays themselves, though not ideal, are also not abject failures. Especially interesting is the apparently straightforward use of TEV protease to cleave surface-displayed proteins. To my knowledge, the first such use of TEV protease in a bacterial display system.

Future Directions

I believe that the TEV-cleavage based detection of the cyclization of TEV-Cys-β2β3-Cys-TEV shows promise. Though it did not generate enough signal here, some of the peaks in the spectrum (Figure B.4) could well be our peptide of interest. It is possible that at higher concentration, and with more specialized instrumentation, the cyclized and non-cyclized peptides would be visible and distinct.

Further, one can imagine interesting assays in which a binding event obscures the TEV site. With such a system, the target of binding would not need to be fluorescently labelled. Instead, a pre-incubated sample could be subjected to TEV cleavage (where TEV cleavage would remove the *myc* tag used to measure display) and washed. Following this wash, an incubation with FITC conjugated anti-*myc* antibody would highlight only those cells which had been binding the target. A flow chart is shown in Figure B.7

The maleimide-FITC conjugations represent a different sort of challenge, since the non-specific binding/uptake seem inherent. My best proposal for moving this project avenue forward is to titrate the maleimide FITC, rather than adding it all at once, but I have no real hope that that would actually solve the non-specificity problem. It is possible that some specific kind of reaction conditions could provide the necessary selectivity, but I have no particular suggestions about what those could be.

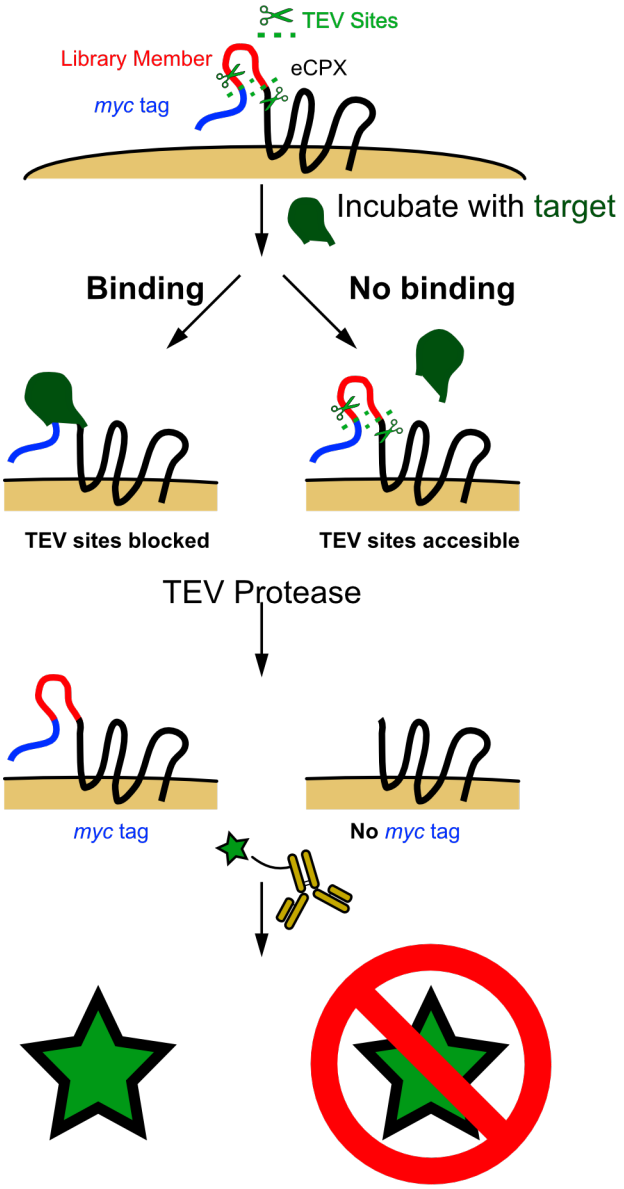


Figure B.7: Proposed use of TEV Cleavage on a Bacterial Surface to Detect Binding Events The protocol proposed here could allow flow cytometry screening without the necessity for the "bait" protein to have a fluorophore attached. Additionally, it could allow some level of kinetic control and detection—intuitively, a fast on/fast off interaction will have more TEV cleavage activity than a slow on/slow off interaction.

B.3 TBP 6.7 Expresses in Mammalian Cells

B.3.1 Introduction

While my *in vitro* work with TBP 6.7 and TAR was enlightening, any real pharmaceutical conclusions can only be drawn from data gathered from *in cellulo* settings. Though I was not responsible for the full experimental work, and it has yet to be published, I did prepare the constructs to be used and do some basic due diligence to make sure they expressed.

B.3.2 Materials Preparation

Cloning

TBP 6.7 sequence was codon-optimized for expression in mammalian cells, and the sequence ordered as a gBlock from IDT. This sequence was PCR amplified to add three different terminal possiblities, a C-terminal FLAG Tag, a C-terminal Nuclear Localization Site (NLS) (specifically the NLS from residues 155–170 of nucleoplasmin [194]), or simply a stop codon following the TBP 6.7 sequence. Primers were designed to give each PCR amplicon the appropriate ending sequence: a stop codon alone (TBP 6.7-*), a FLAG tag (TBP 6.7-FLAG), or the selected NLS (TBP 6.7-NLS). A pcDNA3.0 vector was prepared by digesting with BamHI and XbaI followed by CIP treatment. Amplicons were digested with BsaI, which had been designed to contain the appropriate overhangs to ligate with the cut pcDNA3.0 vector.

Primer and gBlock, as well as DNA and protein sequences can be found in Sections C.6.5-C.6.9.

Cell Culture

HEK 293T cells were prepared according to standard mammalian cell culture protocol in Dulbecco's Modified Eagle Medium (DMEM) with 10% Fetal Bovine Serum in 75 cm^2 Corning flasks. HEK cells were transfected with 3 μg of the TBP 6.7-FLAG construct, using a Lipofectamine 2000 kit (Thermo), with a separate untransfected culture kept as a control.

Western Blot

HEK 293T cells were resuspended in 1 mL and lysed. 2 or 10 µL of lysate was analyzed via Western Blot with an anti-FLAG primary antibody, and a fluorescently labelled secondary antibody. Blots were imaged with a Typhoon imager. The 10 µL samples were too concentrated, and did not run properly via PAGE, but the 2 µL samples ran cleanly. Though there was the non-specific signal associated with use of a FLAG tag in mammalian cells, there was a clear band at the proper molecular weight for TBP 6.7 (~15 kDa) in the transfected cells, but not in the untransfected cells. The blot can be seen in Figure B.8.

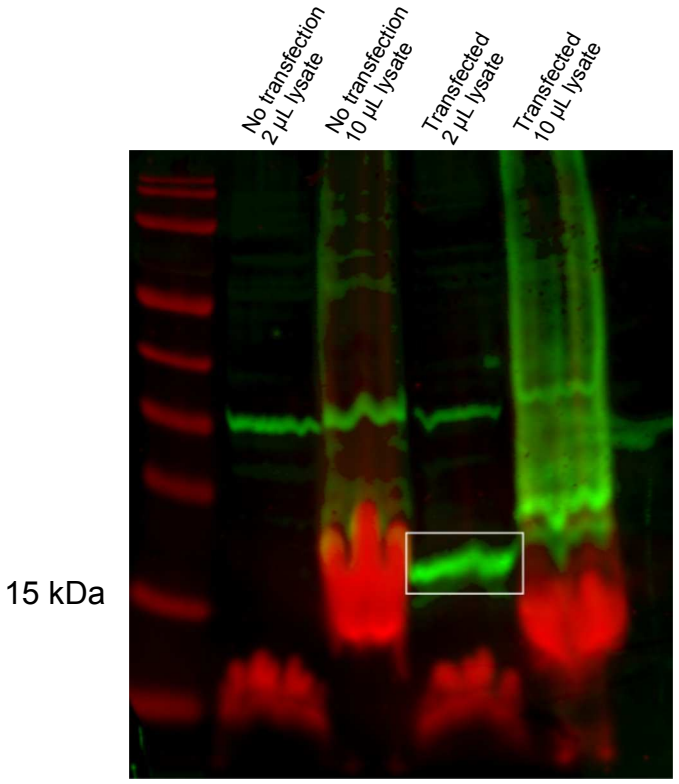


Figure B.8: Western Blot Demonstrating TBP 6.7 Expression in HEK 293T Cells This Western Blot shows that TBP 6.7 expresses cleanly in HEK 293T cells, without apparent toxicity, but only in cells transfected with the appropriate plasmid.

B.3.3 Conclusions

Though this section doesn't represent any data regarding the properties of TBP 6.7 binding to TAR, or preventing HIV infection, it *does* demonstrate that TBP 6.7 expresses, folds, and is apparently non-toxic in the context of a mammalian cell.

B.4 Sac7d Based Binders of CUG_{10} RNA

My initial exposure to Sac7d-based yeast display and libraries was in the context of finding potential α-helical binders to CUG_{10} RNA. This project was primarily the purview of my labmate Rachel Tennyson, with the assistance of our labmate Patrick Beardslee, but as I was deeply involved in both experimental design and assay performance, especially of the yeast display portion of the project, I would like to include a short summary here.

B.4.1 Introduction

CUG repeat RNA is a tempting target, since mis-spliced CUG-repeat RNA is largely responsible for certain forms of muscular dystrophy [29]. As a proxy for expanded CUG repeats, we used CUG_{10} RNA (5′-CCG CUG CUG CUG CUG CUG CUG CUG CUG CUG CUG CGG-3′), with a 5′ cyanine-5 fluorescent dye to measure binding.

We designed our library to be used with yeast display. The primary difference between this yeast display system and the other systems discussed in this thesis (i.e. Sections 2.4 and A.5) is that the *myc* tag was located on the N-terminal of the displayed protein. This is due to the fact that we did not want interference between the *myc* tag and the possible helical binding of CUG_{10} RNA.

B.4.2 Library Creation and Screening

The library was created by randomizing 5 positions on the Sac7d helix (theoretical library size of 3.2×10^6 protein sequences). Library screening was performed in a similar fashion to that described in Section 2.6, with Cyanine-5 labelled CUG_{10} as the target and *E. coli* tRNAs as the off-

target competitor. Though there was no diversification step, the general idea was the same, with each round decreasing CUG_{10} RNA concentration and/or increasing *E. coli* tRNA concentration. The Sac7d based library was screened against other RNAs as well, but the CUG_{10} RNA library had the most promise, and was therefore analyzed most extensively.

B.4.3 Yeast Display Methods

Generally speaking, yeast display for these experiments was conducted in a similar fashion to the many other yeast display experiments throughout this thesis (notably Section 2.4.2).

Briefly, samples which involved FITC-conjugated anti-*myc* antibody were incubated with 1:1000 antibody. All yeast display samples also contained a 1:2000 dilution of rRNasin (Promega). All RNA underwent a standard melt and refold (Section 2.4.2) prior to use in yeast display. Incubations were 45 minutes to 1 hour, in 200–1000 µL sample volumes with PBS-BSA as the buffer.

B.4.4 Analysis of the Best Binder: CUG_{10} Binding Protein 5.21

Antibody vs. CUG_{10} RNA Binding

The apparent best binder of RNA was $\underline{C}UG_{10}$ $\underline{B}$inding $\underline{P}$rotein 5.21 (CBP 5.21). There were some issues characterizing this protein. One prominent concern is the fact that both antibody and CUG_{10} RNA binding seemed to occur reasonable robustly separately, but not concurrently. Graphically, what this looks like is robust positive signals for both the FITC channel (representing binding to FITC conjugated anti-*myc* antibody, generally the X-axis) *or* the Cy-5 channel (representing CUG_{10} RNA binding, generally the Y-axis), but with minimal signal in the "double positive" quadrant (data shown in Figure B.9).

CBP 5.21 Binding to CUG_{10} RNA

Ultimately, comparing the apparent ability to bind CUG_{10} RNA of CBP 5.21 and the original Sac7d upon which the library was based makes CBP 5.21 seem like a promising lead, shown by the concentration series in Figure B.10. In this experiment, yeast displaying either Sac7d or CBP 5.21 were incubated with varying concentrations of CUG_{10} RNA ranging from 1–10 µM. Worth not-

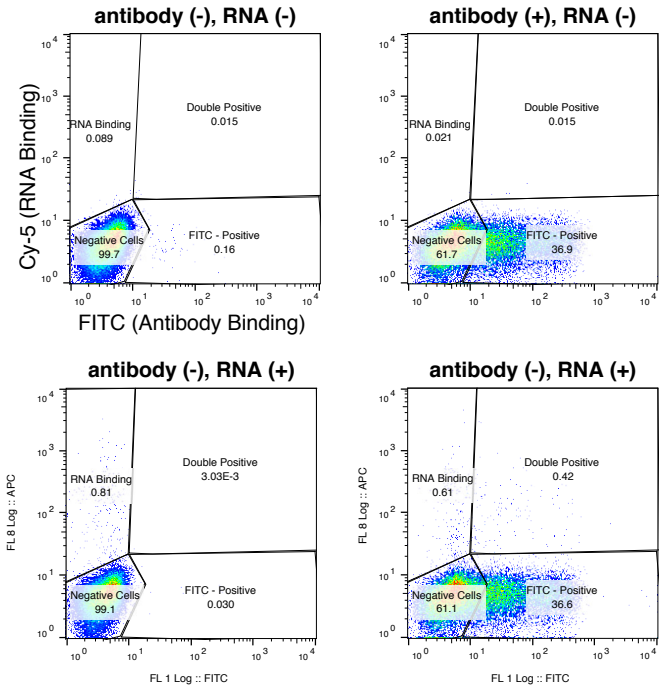


Figure B.9: CBP 5.21 Binding to Cy-5 Labelled CUG_{10} RNA or FITC conjugated anti-*myc* Antibody CBP 5.21 binds to antibody *or* CUG_{10} RNA, but does not display the expected behavior of binding antibody *and* CUG_{10} RNA.

ing is that these concentrations are quite high, and simple non-specific binding increased the fluorescence of the negative population appreciably. Gating was adjusted throughout the experiment in order to exclude this negative population.

B.4.5 Conclusions

Though we never were able to fully demonstrate CUG_{10} binding via an *in vitro* method (such as ITC or ELISA), the yeast display data is fairly promising. The Sac7d-displaying yeast always exhibit a similar level of non-specific binding no matter the concentration, while the CBP 5.21-displaying yeast bind *more* RNA in absolute terms (both percent positive and with higher signal), and exhibit concentration dependence.

There are, of course, concerns. The most notable issue is the difficulty in *correlating* display and binding. One possible explanation is that the displayed protein is not being displayed at all, and is simply being cleaved off at the *myc* tag. If this is the case, our FITC-conjugated anti-*myc* antibody is misleading us, and is showing mostly cells that do *not* have displayed protein. The

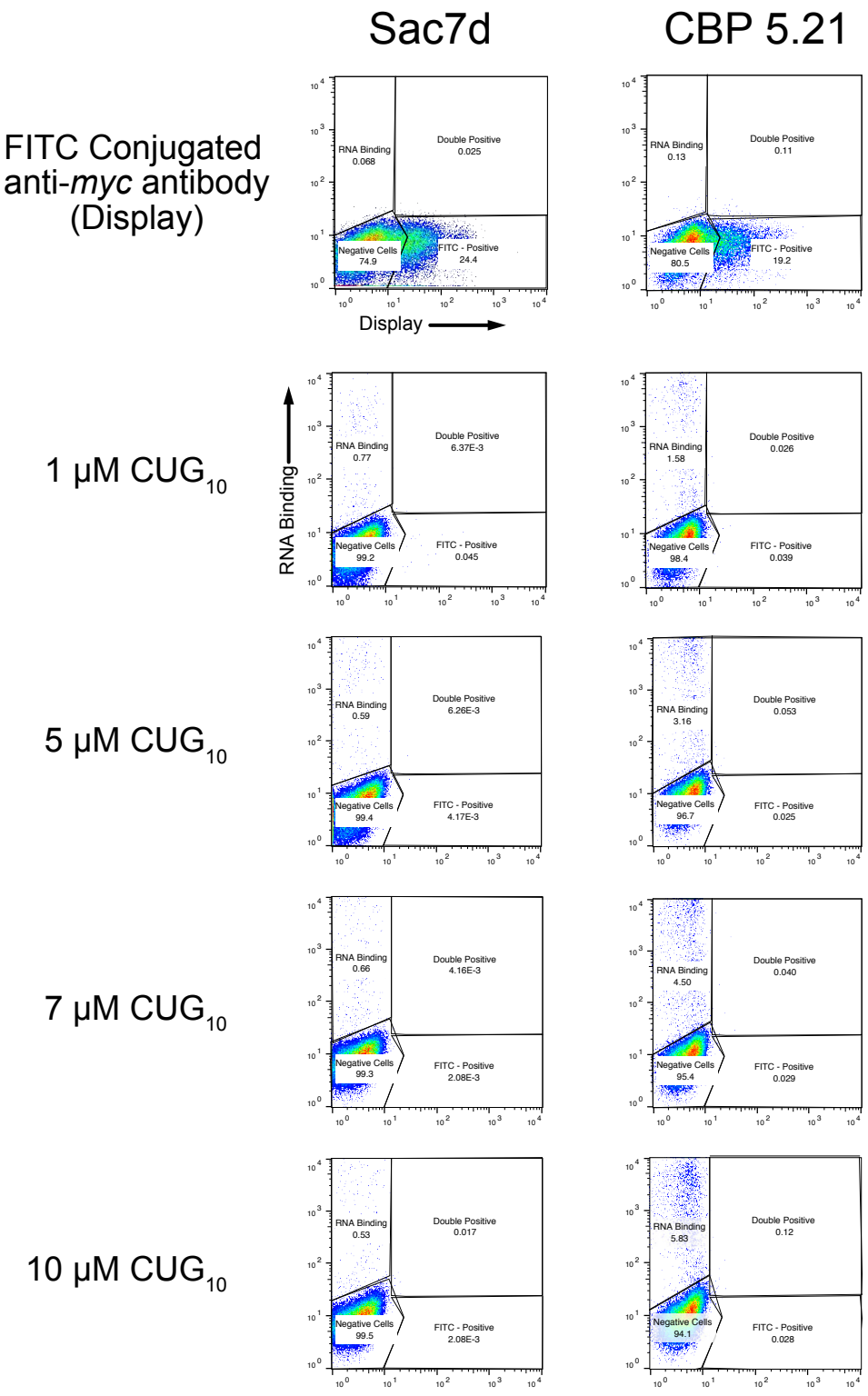


Figure B.10: Concentration Series of Displayed CBP 5.21 Incubated with 1–10 μM Cy5-CUG_{10}RNA] This concentration series indicates that our CBP 5.21 protein binds CUG_{10}RNA in the context of yeast display. The percent of the population in each quadrant is indicated on the respective graph.

only way around this is to add the display tag to the C-terminal, which could interfere with the essential helical nature of this region.

Another possibility is that there is a steric problem with antibody binding between the N-terminal of the Sac7d protein and the Aga2/yeast surface. In this case, antibody-binding would favor (and possibly force) odd conformations and folding patterns which may create a contradiction between the Sac7d variant's ability to bind antibody and RNA. In this case, a possible way forward is to use a smaller tag/detector combo, for instance a BC2 tag and BC2 nanobody pair, with a fluorescent tag on the BC2 nanobody [195, 196].

Ultimately, though it would be desirable to have more clarity, the data in Figure B.10 demonstrates a clear concentration-dependent increase in CUG_{10} 10 RNA binding which is dependent upon the presence of the Sac7d variant CBP 5.21, which can only be seen as a promising sign.

B.5 Others

Finally, I'd like to give brief mention of a few additional projects

B.5.1 Enzymatic Creation of Inorganic Nanoparticles

My first publication was not, in fact, in the McNaughton Lab, but came from work done during my initial six-week rotation in the lab of Prof. Chris Ackerson [197]. There was a developing project relating to the ability of *Pseudomonas moraviensis stanleyae* to high levels of selenite, possibly reducing the selenite—$SeO_3)^{-2}$—into insoluble, neutral selenium (Se). It was suggested that the enzyme responsible was glutathione reductase, and I discovered that it was possible to use commercially available glutathione reductase to perform the transformation. I also hit upon the idea of monitoring the reaction via NADPH, which absorbs at 340 nM (spectra shown in Figure B.11).

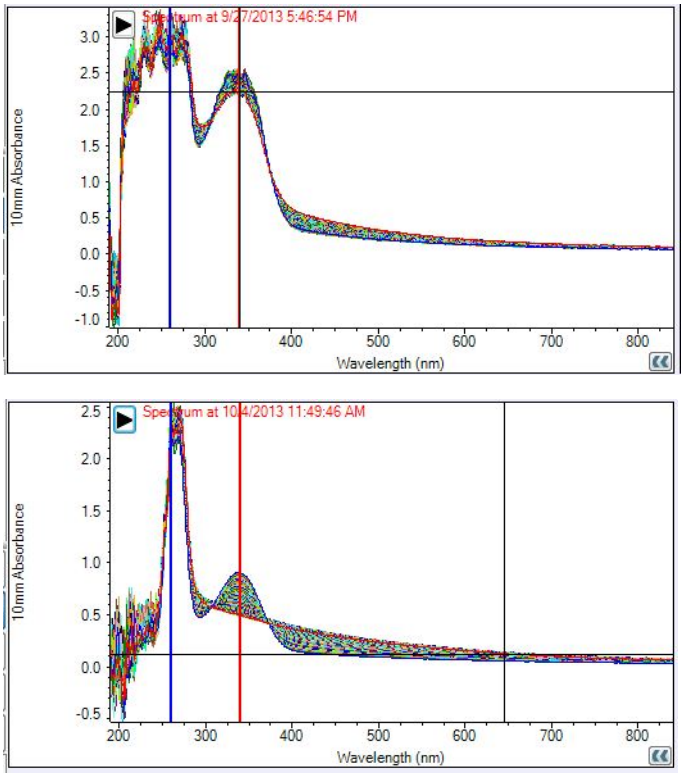


Figure B.11: NADPH-based Monitoring of Enzymatic Selenite Reduction My first experimental success in graduate school was in determining a protocol for measuring the rate (and therefore allowing condition optimization) of glutathione reductase based reduction of selenite to neutral selenium.

B.5.2 Alternate Library Selection Method

I probably spent as much time working on an original idea of mine involving a proposed general method to evaluate and screen a library based on *selection* rather than flow cytometry based *screening* as I did on everything described in Chapters 2-5. Publication forthcoming.

B.5.3 Small Molecule Induced Dimerization

I also worked closely with graduate student colleague Patrick Beardslee and undergraduate researcher Zachary Fleishhacker on a project involving the well-known FRB–FKBP–rapamycin interaction [198]. FRB and FKBP are proteins which only interact when small molecule rapamycin is present. We worked to build a bacterial-display platform to study this interaction, but had no success. Between the work in this section and the prior section, I would guess that my experience with bacterial display is even more extensive than the significant yeast display work described throughout this thesis.

B.5.4 Library Screening for Other RNAs

I worked closely with graduate student colleague Angeline Ta, and post-doc Gayani Perera to retreat the steps in Chapter 2 to find binders for the VP30 RNA, and miR-21. Though we built new and useful U1A-based library creation platforms, we had no success in finding binders for additional RNAs.

Appendix C

Protein and DNA Sequences

C.1 Sequences from Chapter 2, "Affinity Maturation of U1A E19S for TAR RNA Binding"

C.1.1 Selected Primers from Chapter 2

Fwd U1A NheI

5′-ATA TAG CTA GCA TGG CCC AGG TGC AGC-3′

Rev U1A BamHI

5′-CGG GAT CCT GCG GCC GCA ACC-3′

Fwd BsaI Out

5′-CAA GGA GGT GTC GAG C GCC ACC AAC-3′

Rev BsaI Out

5′-CTC GAC ACC TCC TTG AAG ATG ACA AAA GCT TGG CC-3′

FWD U1A BsaI

5′-ATA TAG CTA GCA GCT AGC TAG CTA GAT GGT CTC AGG GGC CAA GCT TTT GTC ATC TTC AAG GAG GTT TCG-3′

Rev b2- b3 lib

5′-TAT AT GGT CTC GCC CCT MNN MNN MNN MNN CCG MNN TAC CAG GAT ATC CAG GAT CTG GCC-3′

HR FP

5′-CTC TGG TGG AGG GCG TAG CGG AGG CGG AGG GTC GGC TAG C-3′

HR RP

5′-CGA GCT ATT ACA AGT CCT CTT CAG AAA TCA GCT TTT GTT CGG ATC C-3′

Fwd c-helix receiving

5′-CGC GTC CTA ACC ACA CTA TTT ATA TGA GAC CAC TCT AGA GGT TCC CCG GTT GC-3′

Rev c-helix receiving

5′-GGC CGC AAC CGG GGA ACC TCT AGA GTG GTC TCA TAT AAA TAG TGT GGT TAG GA-3′)

Rev c-helix lib

5′-TAT ATG GTC TCT TGG CMN NGA TMN NMN NGT CGG TGC GCG CAT ACT GGA TAC G-3′

Fwd b1a1

5′-ATA TAG GTC TCT TAT ATC NNK NNK CTC AAT NNK AAG ATC AAG AAG GAT GAG CTC AAA AAG-3′

C.1.2 wtU1A in pCTcon2

Protein Sequence

DNA Sequence

The text in bold corresponds to the NheI and BamHI cut sites. The DNA sequence of all pCTcon2 fusion proteins listed for the remainder of this thesis will be shown beginning at the NheI site and ending with the BamHI site.

5′-ATGCAAGGAGTTTTTGAATATTACAAATCAGTAACGTTTGTCAGTAATTGCGGTT CTCACCCCTCAACAACTAGCAAAGGCAGCCCCATAAACACACAGTATGTTTTTAAGG ACAATAGCTCGACGATTGAAGGTAGATACCCATACGACGTTCCAGANTACGCTCTGC

AGGCTAGTGGTGGAGGAGGCTCTGGTGGAGGCGGTAGCGGAGGCGGAGGGTCG**GC-**

TAGCATGGCCCAGGTGCAGCTGCAGGTCGACATGGCAGTTCCCGAGACGCGTCCTAA

CCACACTATTTATATCAACAACCTCAATGAGAAGATCAAGAAGGATGAGCTCAAAAA

GTCCCTGTACGCCATCTTCTCCCAGTTTGGCCAGATCCTGGATATCCTGGTATCACG

GAGCCTGAAGATGAGGGGCCAAGCTTTTGTCATCTTCAAGGAGGTCTCGAGCGCCAC

CAACGCCCTGCGCTCCATGCAGGGTTACCCTTTCTATGACAAACCTATGCGTATCCA

GTATGCGCGCACCGACTCAGATATCATTGCCAAGATGAAAGGCACCTTCGGATCGGT

CGACTCTAGAGGTTCCCCGGTTGCGGCCGCA**GGATCC**GAACAAAAGCTTATTTCTGA

AGAGGACTTGTAATAGCTCG-3′

C.1.3 U1A E19S in pCTcon2

Protein Sequence

E19S mutation in **bold**

N-ASMAQVQLQVDMAVPETRPNHTIYINNLN**S**KIKKDELKKSLYAIFSQFGQILDILVSRSLK

MRGQAFVIFKEVSSATNALRSMQGYPFYDKPMRIQYARTDSDIIAKMKGTFGSVDSRGS-C

DNA Sequence

NheI and BamHI cloning sites in **bold**

5′-**GCTAGC**ATGGCCCAGGTGCAGCTGCAGGTCGACATGGCAGTTCCCGAGACGCGTCC

TAACCACACTATTTATATCAACAACCTCAATGAGAAGATCAAGAAGGATGAGCTCAA

AAAGTCCCTGTACGCCATCTTCTCCCAGTTTGGCCAGATCCTGGATATCCTGGTATCA

CGGAGCCTGAAGATGAGGGGCCAAGCTTTTGTCATCTTCAAGGAGGTCTCGAGCGCC

ACCAACGCCCTGCGCTCCATGCAGGGTTACCCTTTCTATGACAAACCTATGCGTATC

CAGTATGCGCGCACCGACTCAGATATCATTGCCAAGATGAAAGGCACCTTCGGATCG

GTCGACTCTAGAGGTTCCCCGGTTGCGGCCGCA**GGATCC**-3′

C.1.4 1st Gen Library Receiving Plasmid/BsaI U1A

NheI and BamHI in **bold**, BsaI site in *italics*, //overhang//

5′-**GCTAGC**TAGAT*GGTCTCA*//GGGG//CCAAGCTTTTGTCATCTTCAAGGAGGTTTCGA
GCGCCACCAACGCCCTGCGCTCCATGCAGGGTTACCCTTTCTATGACAAACCTATGC
GTATCCAGTATGCGCGCACCGACTCAGATATCATTGCCAAGATGAAAGGCACCTTCG
GATCGGTCGACTCTAGAGGTTCCCCGGTTGCGGCCGCA**GGATCC** -3′

C.1.5 Library Amplicon

Homologous recombination regions in **bold**

5′-**CTCTGGTGGAGGCGGTAGCGGAGGCGGAGGGTCGGCTAGC**ATGGCCCAGGTGCAGC
TGCAGGTCGACATGGCAGTTCCCGAGACGCGTCCTAACCACACTATTTATATCAACA
ACCTCAATGAGAAGATCAAGAAGGATGAGCTCAAAAAGTCCCTGTACGCCATCTTCT
CCCAGTTTGGCCAGATCCTGGATATCCTGGTANNKCGGNNKNNKNNKNNKAGGGGCC
AAGCTTTTGTCATCTTCAAGGAGGTCTCGAGCGCCACCAACGCCCTGCGCTCCATGC
AGGGTTACCCTTTCTATGACAAACCTATGCGTATCCAGTATGCGCGCACCGACTCAG
ATATCATTGCCAAGATGAAAGGCACCTTCGGATCGGTCGACTCTAGAGGTTCCCCGG
TTGCGGCCGCA**GGATCCGAACAAAAGCTTATTTCTGAAGAGGACTTGTAATAGCTCG**-3′

C.1.6 2nd Gen Library Receiving Plasmid

NheI and BamHI in **bold**, BsaI site in //*italics*//

5′-**GCTAGC**ATGGCCCAGGTGCAGCTGCAGGTCGACATGGCAGTTCCCGAGACGCGT
CCTAACCACACTATTTATAT//*GAGACC*//ACTCTAGAGGTTCCCCGGTTGCGGCCGCA
GGATCC-3′

C.1.7 2nd Gen Library Amplicon

Homologous recombination regions in **bold** β2β3 random bases denoted with ”X” 5′-**CTCTG-**

GTGGAGGCGGTAGCGGAGGCGGAGGGTCGGCTAGCATGGCCCAGGTGCAGCTGCAG
GTCGACATGGCAGTTCCCGAGACGCGTCCTAACCACACTATTTATATCAACAAC
CTCAATGAGAAGATCAAGAAGGATGAGCTCAAAAAGTCCCTGTACGCCATCTTC

TCCCAGTTTGGCCAGATCCTGGATATCCTGGTAXXXCGGXXXXXXXXXXXXAGGGG
CCAAGCTTTTGTCATCTTCAAGGAGGTCTCGAGCGCCACCAACGCCCTGCGCTC
CATGCAGGGTTACCCTTTCTATGACAAACCTATGCGTATCCAGTATGCGCGCAC
CGACNNKGATNNKNNKGCCAAGATGAAAGGCACCTTCGGATCGGTCGACTCTAGA
GGTTCCCCGGTTGCGGCCGCA**GGATCCGAACAAAAGCTTATTTCTGAAGAGGACTTG-**
TAATAGCTCG-3′

C.2 Sequences from Chapter 3, "Characterization of TAR Binding Proteins"

C.2.1 Selected Primers from Chapter 3

U1A NcoI FP

5′-ATA TAC CAT GGC CCA GGT GCA GC-3′

U1A His FLAG PacI RP

5′-GTT AAT TAA CTA TTA CTT GTC GTC ATC GTC TTT GTA GTC GTG ATG ATG GTG ATG ATG TGC GGC CGC AAC C-3′

U1A His PacI RP

5′-GGT TGC GGC CGC ACA TCA TCA CCA TCA TCA CTA ATA GTT AAT TAA C-3′

PLAI FP

5′-TCTAGAACTAGTGGATCTTAG-3′

PLAI RP

5′-GCTACAACCATCCCTTCAGAC-3′

C.2.2 Generic TAR Binding Protein with C-terminal His_6 and FLAG Tags

Protein Sequence

Variable positions in β2β3 loop or C-helix designated with [#], where # is the position number.

Canonical position 1 designated with "***M***".

N-MAQVQLQVD***M***AVPETRPNHTIYINNLNSKIKKDELKKSLYAIFSQFGQILDILV[47]R[49][50][51][52]RGQAFVIFKEVSSATNALRSMQGYPFYDKPMRIQYARTD[91][92]I[94]AKMKGTFGSVDSRGSPVAAAHHHHHHVAADYKDDDDK-C

DNA Sequence

ORF denoted by "||".

β2β3 loop variable bases denoted by X, C-helix variable bases with Y.

NcoI and PacI cloning sites indicated by **bold**.

5′-**CC**||***ATG*****G**CCCAGGTGCAGCTGCAGGTCGACATGGCAGTTCCCGAGACGCGTCCTAACCACACTATTTATATCAACAACCTCAATTCGAAGATCAAGAAGGATGAGCTCAAAAAGTCCCTGTACGCCATCTTCTCCCAGTTTGGCCAGATCCTGGATATCCTGGTAXXXCGGXXXXXXXXXXXXAGGGGCCAAGCTTTTGTCATCTTCAAGGAGGTTTCGAGCGCCACCAACGCCCTGCGCTCCATGCAGGGTTACCCTTTCTATGACAAACCTATGCGTATCCAGTATGCGCGCACCGACNNKNNKATCNNKGCCAAGATGAAAGGCACCTTCGGATCGGTCGACTCTAGAGGTTCCCCGGTTGCGGCCGCACATCATCACCATCATCACGTGGCCGCAGACTACAAAGACGATGACGACAAG||TAATAG**TTAATTAA**-3′

C.2.3 TBP 6.7

Protein Sequence of TBP 6.7 with C-term His_6 and FLAG Tags

Canonical position 1 designated with "***M***".

N-MAQVQLQVD***M***AVPETRPNHTIYINNLNSKIKKDELKKSLYAIFSQFGQILDILVPRQRTPRGQAFVIFKEVSSATNALRSMQGYPFYDKPMRIQYARTDSDIIAKMKGTFGSVDSRGSPVAAAHHHHHHDYKDDDDK-C

DNA Sequence of TBP 6.7 with C-terminal His_6 and FLAG Tags

NcoI and PacI cloning sites indicated by **bold.**

5′-**CC**||***ATG*****G**CCCAGGTGCAGCTGCAGGTCGACATGGCAGTTCCCGAGACGCGTCCTAA
CCACACTATTTATATCAACAACCTCAATTCGAAGATCAAGAAGGATGAGCTCAAAAA
GTCCCTGTACGCCATCTTCTCCCAGTTTGGCCAGATCCTGGATATCCTGGTACCGCG
GCAGCGGACGCCGAGGGGCCAAGCTTTTGTCATCTTCAAGGAGGTTTCGAGCGCCAC
CAACGCCCTGCGCTCCATGCAGGGTTACCCTTTCTATGACAAACCTATGCGTATCCA
GTATGCGCGCACCGACAAGCGTATCCCGGCCAAGATGAAAGGCACCTTCGGATCGGT
CGACTCTAGAGGTTCCCCGGTTGCGGCCGCACATCATCACCATCATCACGTGGCCGC
AGACTACAAAGACGATGACGACAAG||TAATAG**TTAATTAA**-3′

Protein Sequence of TBP 6.7 with C-terminal His_6 Tag

Canonical position 1 designated with "***M***".

N-MAQVQLQVD***M***AVPETRPNHTIYINNLNSKIKKDELKKSLYAIFSQFGQILDILVPRQRTP
RGQAFVIFKEVSSATNALRSMQGYPFYDKPMRIQYARTDSDIIAKMKGTFGSVDSRGSPVA
AAHHHHHH-C

C.2.4 TBP 6.6

Protein Sequence of TBP 6.6 with C-terminal His_6 and FLAG Tags

Canonical position 1 designated with "***M***".

N-MAQVQLQVD***M***AVPETRPNHTIYINNLNSKIKKDELKKSLYAIFSQFGQILDILVPRTRTP
RGQAFVIFKEVSSATNALRSMQGYPFYDKPMRIQYARTDSDIIAKMKGTFGSVDSRGSPVA
AAHHHHHHVAADYKDDDDDK-C

DNA Sequence of TBP 6.6 with C-terminal His_6 and FLAG Tags

NcoI and PacI cloning sites indicated by **bold.**

5′-**CC**||***ATGG***CCCAGGTGCAGCTGCAGGTCGACATGGCAGTTCCCGAGACGCGTCCTAA
CCACACTATTTATATCAACAACCTCAATTCGAAGATCAAGAAGGATGAGCTCAAAAA
GTCCCTGTACGCCATCTTCTCCCAGTTTGGCCAGATCCTGGATATCCTGGTACCGCG
GACGCGGACTCCGAGGGGCCAAGCTTTTGTCATCTTCAAGGAGGTTTCGAGCGCCAC
CAACGCCCTGCGCTCCATGCAGGGTTACCCTTTCTATGACAAACCTATGCGTATCCA
GTATGCGCGCACCGACGGGAGGATCGCGGCCAAGATGAAAGGCACCTTCGGATCGGT
CGACTCTAGAGGTTCCCCGGTTGCGGCCGCACATCATCACCATCATCACGTGGCCGC
AGACTACAAAGACGATGACGACAAG||TAATAG**TTAATTAA**-3′

Protein Sequence of TBP 6.6 with C-terminal His_6 Tag

Canonical position 1 designated with "***M***".

N-MAQVQLQVD***M***AVPETRPNHTIYINNLNSKIKKDELKKSLYAIFSQFGQILDILVPRTRTP
RGQAFVIFKEVSSATNALRSMQGYPFYDKPMRIQYARTDGRIAAKMKGTFGSVDSRGSPV
AAAHHHHHH-C

C.2.5 RNAs

TAR

The TAR RNA used in the ITC assay shown in Figure 3.15 did not have a biotin modification.

5′-biotin GGC AGA UCU GAG CCU GGG AGC UCU CUG CC-3′

U1hpII

5′-biotin AGC UUA UCC AUU GCA CUC CGG AUG AGC-3′

C.2.6 Tat Sequences

Tat peptide (Used in ITC)

N-RKKRRQRRRPPQGSQTHQVSLSKQPTSQPRGDPTGPKE-C

Tat Protein from Prospec (Used in Transcription Assay)

ProSpec Recombinant HIV-1 TAT Clade-B (Cat. No. HIV-129)

N-MEPVDPRLEPWKHPGSQPKTACTNCYCKKCCFHCQVCFITKALGlSYGRKKRRQRRRPP
QGSQTHQVSLSKQPTSQSRGDPTGPKE-C

C.2.7 PLAI-BS Transcript Sequence

Primers used for amplification **bolded**, TAR element in *italics*.

5′-**TCTAGAACTAGTGGATCTTAG**CCACTTTTTAAAAGAAAAGGGGGGACTGGAAGGGC
TAATTCACTCCCAACGAAGACAAGATATCCTTGATCTGTGGATCTACCACACACAAG
GCTACTTCCCTGATTGGCAGAACTACACACCAGGGCCAGGGGTCAGATATCCACTG
ACCTTTGGATGGTGCTACAAGCTAGTACCAGTTGAGCCAGATAAGGTAGAAGAGGC
CAATAAAGGAGAGAACACCAGCTTGTTACACCCTGTGAGCCTGCATGGAATGGATG
ACCCTGAGAGAGAAGTGTTAGAGTGGAGGTTTGACAGCCGCCTAGCATTTCATCAC
GTGGCCCGAGAGCTGCATCCGGAGTACTTCAAGAACTGCTGACATCGAGCTTGCTA
CAAGGGACTTTCCGCTGGGGACTTTCCAGGGAGGCGTGGCCTGGGCGGGACTGGGG
AGTGGCGAGCCCTCAGATGCTGCATATAAGCAGCTGCTTTTTGCCTGTACTGGGTC
TCTCTGGTTAGA*CCAGATTTGAGCCTGGGAGCTCTCTGGC*TAACTAGGGAACCCACTGC
TTAAGCCTCAATAAAGCTTGCCTTGAGTGCTTCAAGTAGTGTGTGCCCGTCTGTTGT
GTGACTCTGGTAACTAGAGATCCCTCAGACCCTTTTAGTCAGTGTGGAAAATCTCTA
GCAGTGGCGCCCGAACAGGGACTTGAAAGCGAAAGGGAAACCAGA**GGAGCTCTCTC-**
GACGCAGGACT-3′

C.3 Sequences from Chapter 4, "Crystallization of TBP 6.7 and TAR"

C.3.1 TBP 6.7 Variants

Full-Length TBP 6.7

Truncations indicated with "//"

N-MAQVQLQVD//MAVPETRPNHTIYINNLNSKIKKDELKKSLYAIFSQFGQILDILVPRQRT
PRGQAFVIFKEVSSATNALRSMQGYPFYDKPMRIQYARTDKRIPAKMK//GTFGSVDSRGSP
VAAAHHHHHHDYKDDDDK-C

Truncated TBP 6.7

N-MAVPETRPNHTIYINNLNSKIKKDELKKSLYAIFSQFGQILDILVPRQRTPRGQAFVIFKEV
SSATNALRSMQGYPFYDKPMRIQYARTDKRIPAKMKAHHHHHHDYKDDDDK-C

TBP 6.7 for Crystallography

TEV cleavage site indicated by \\. Protein was only used in crystallography following removal of this tag using TEV protease.

N-MGSSHHHHHHSSGENLYFQ\\GHMAVPETRPNHTIYINNLNSKIKKDELKKSLHAIFSRF
GQILDILVPRQRTPRGQAFVIFKEVSSATNALRSMQGFPFYDKPMRIQYAKTDKRIPAKMK
-C

TBP 6.7 Y75F

F75 in **bold**

N-MAVPETRPNHTIYINNLNSKIKKDELKKSLYAIFSQFGQILDILVPRQRTPRGQAFVIFKEV
SSATNALRSMQG**F**PFYDKPMRIQYARTDKRIPAKMKAHHHHHHDYKDDDDK-C

TBP 6.7 R88K

R88 in **bold**

N-MAVPETRPNHTIYINNLNSKIKKDELKKSLYAIFSQFGQILDILVPRQRTPRGQAFVIFKEV
SSATNALRSMQGYPFYDKPMRIQYA**K**TDKRIPAKMKAHHHHHHDYKDDDDK-C

TBP 6.7 Y75F/R88K

F75 and R88 in **bold**

N-MAVPETRPNHTIYINNLNSKIKKDELKKSLYAIFSQFGQILDILVPRQRTPRGQAFVIFKEV
SSATNALRSMQG**F**PFYDKPMRIQYA**K**TDKRIPAKMKAHHHHHHDYKDDDDK-C

C.4 Sequences from Chapter 5, "Peptide Derivatives of TBP 6.7"

C.4.1 TBP 6.7 Used in ITC Assays

TEV cleavage position indicated by //. Protein was only used in crystallography following removal of this TEV cleavage site

N-MGSSHHHHHHSSGENLYFQ//GHMAVPETRPNHTIYINNLNSKIKKDELKKSLHAIFSRF
GQILDILVPRQRTPRGQAFVIFKEVSSATNALRSMQGFPFYDKPMRIQYAKTDKRIPAKMK
-C

C.4.2 SUMO Fusions

SUMO Control

N-MDYKDDDDKHHHHHHMSDSEVNQEAKPEVKPEVKPETHINLKVSDGSSEIFFKIKKTT
PLRRLMEAFAKRQGKEMDSLRFLYDGIRIQADQTPEDLDMEDNDIIEAHREQIGGATY-C

SUMO β2β3

N-MDYKDDDDKHHHHHHMSDSEVNQEAKPEVKPEVKPETHINLKVSDGSSEIFFKIKKTT
PLRRLMEAFAKRQGKEMDSLRFLYDGIRIQADQTPEDLDMEDNDIIEAHREQIGGATYGG
GGSLDILVPRQRTPRGQAFVIF-C

SUMO β2β3 (R47A)

N-MDYKDDDDKHHHHHHMSDSEVNQEAKPEVKPEVKPETHINLKVSDGSSEIFFKIKKTT
PLRRLMEAFAKRQGKEMDSLRFLYDGIRIQADQTPEDLDMEDNDIIEAHREQIGGATYGG
GGSLDILVPAQRTPRGQAFVIF-C

SUMO β2β3 (R49A)

N-MDYKDDDDKHHHHHHMSDSEVNQEAKPEVKPEVKPETHINLKVSDGSSEIFFKIKKTT
PLRRLMEAFAKRQGKEMDSLRFLYDGIRIQADQTPEDLDMEDNDIIEAHREQIGGATYGG
GGSLDILVPRQATPRGQAFVIF-C

TBP 6.7-FLAG (Used in ELISA)

N-MAQVQLQVDMAVPETRPNHTIYINNLNSKIKKDELKKSLYAIFSQFGQILDILVPRQRTP
RGQAFVIFKEVSSATNALRSMQGYPFYDKPMRIQYARTDKRIPAKMKGTFGSVDSRGSPV
AAAHHHHHHDYKDDDDK-C

TBP 6.7 (Transcription Assay)

N-MAQVQLQVDMAVPETRPNHTIYINNLNSKIKKDELKKSLYAIFSQFGQILDILVPRQRTP
RGQAFVIFKEVSSATNALRSMQGYPFYDKPMRIQYARTDKRIPAKMKGTFGSVDSRGSPV
AAAHHHHHH-C

C.4.3 Aga2 Control

DNA Sequence

NheI and BamHI cloning sites **bolded**

5′-ATGCAGTTACTTCGCTGTTTTTCAATATTTTCTGTTATTGCTTCAGTTTTAGCACAG
GAACTGACAACTATATGCGAGCAAATCCCCTCACCAACTTTAGAATCGACGCCGTAC
TCTTTGTCAACGACTACTATTTTGGCCAACGGGAAGGCAATGCAAGGAGTTTTTGAA
TATTACAAATCAGTAACGTTTGTCAGTAATTGCGGTTCTCACCCCTCAACAACTAGCA
AAGGCAGCCCCATAAACACACAGTATGTTTTTAAGGACAATAGCTCGACGATTGAAG

GTAGATACCCATACGACGTTCCAGACTACGCTCTGCAGGCTAGTGGTGGAGGAGGCT
CTGGTGGAGGCGGTAGCGGAGGCGGAGGGTCG**GCTAGC**GGAGGCGGAGGGTCGGGA
GGCGGA**GGATCC**GAACAAAAGCTTATTTCTGAAGAGGACTTGT-3′

Protein Sequence

Linker Sequence **bolded**, along with the AS and GS sequences corresponding to the NheI and BamHI sites respectively.

N-MQLLRCFSIFSVIASVLAQELTTICEQIPSPTLESTPYSLSTTTILANGKAMQGVFEYYKSV
TFVSNCGSHPSTTSKGSPINTQYVFKDNSSTIEGRYPYDVPDYALQASGGGGSGGGGSGGG
GS**ASGGGGSGGGGS**EQKLISEEDL-C

C.4.4 TBP 6.7 β2β3 Loop for Yeast Display

DNA Sequence

NheI and BamHI cloning sites **bolded**

N-**GCTAGC**CTGGATATCCTGGTACCGCGGCAGCGGACGCCGAGGGGCCAAGCTTTTGT
CATCTTC**GGATCC**-C

Protein Sequence

The AS and GS sequences at the beginning and end of the sequence correspond to the NheI and BamHI cloning sites.

N-ASLDILVPRQRTPRGQAFVIFGS-C

C.4.5 TBP 6.7 β2β3 Loop–eCPX for Bacterial Display

DNA Sequence

NdeI and XhoI cloning sites in **bold**. Start and end of in-frame β2β3 loop in *italics*.

5′-ATGAAAAAAATTGCATGTCTTTCAGCACTGGCCGCAGTTCTGGCTTTCACCGCA
GGTACTTCCGTAGCTGGTCAGTCTGGCCAGGCGGCCGCTCCCGGGGAACAAAAAC

TGATTTCTGAAGAGGACTTGGGCGCGCCTA**CATATGG**C*CTGGATATCCTGGTACCGCG*
*GCAGCGGACGCCGAGGGGCCAAGCTTTTGTCATCTTC*C**CTCGAG**TCGGTGGCGGAAGCGG
AGGGGGCTCTGGCGGAGGGTCAGGTGGGGGCAGCGGAGGGGGATCGGGAGGGCAGT
CTGGGCAGTCTGGTGACTACAACAAAAACCAGTACTACGGCATCACTGCTGGTCCGG
CTTACCGCATTAACGACTGGGCAAGCATCTACGGTGTAGTGGGTGTGGGTTATGGTA
AATTCCAGACCACTGAATACCCGACCTACAAACACGACACCAGCGACTACGGTTTCT
CCTACGGTGCGGGTCTGCAGTTCAACCCGATGGAAAACGTTGCTCTGGACTTCTCTT
ACGAGCAGAGCCGTATTCGTAGCGTTGACGTAGGCACCTGGATTTTGTCTGTTGGTT
ACCGCTTCGGGAGTAAATCGCGTCGCGCGACTTCTACTGTAACTGGCGGTTACGCAC
AGAGCGACGCTCAGGGCCAAATGAACAAAATGGGCGGTTTCAACCTGAAATACCGCT
ATGAAGAAGACAACAGCCCGCTGGGTGTGATCGGTTCTTTCACTTACACCGAGAAAA
GCCGTACTGCAAGC-3′

Protein Sequence

β2β3Loop sequence **bolded**

N-MKKIACLSALAAVLAFTAGTSVAGQSGQAAAPGEQKLISEEDLGAPTYG**LDILVPRQRT-**
PRGQAFVIFPRVGGGSGGGSGGGSGGGSGGGSGGQSGQSGDYNKNQYYGITAGPAYRIN
DWASIYGVVGVGYGKFQTTEYPTYKHDTSDYGFSYGAGLQFNPMENVALDFSYEQSRIRS
VDVGTWILSVGYRFGSKSRRATSTVTGGYAQSDAQGQMNKMGGFNLKYRYEEDNSPLGV
IGSFTYTEKSRTAS-C

C.5 Sequences from Appendix A, “Helical Grafting of E6AP”

C.5.1 Sac7d for Yeast Display

DNA Sequence

5′-**GCTAGC**CGTGAAAGTGAAATTTCTGCTGAACGGCGAAGAAAAAGAAGTGGATACC
AGCAAAATTCGCGATGTGAGTCGCCAGGGCAAAAACGTGAAATTTACCTATAACGAT

AACGGCAAATATGGCGCGGGCAACGTGGATGAAAAAGATGCGCCGAAAGAACTGCT
GGATATGCTGGCGCGCGCGGAACGCGAAAAAAAACTGAAC**GGATCC**-3′

Protein Sequence

N-**AS**VKVKFLLNGEEKEVDTSKIRDVSRQGKNVKFTYNDNGKYGAGNVDEKDAPKELLD
MLARAEREKKLN**GS**-C

C.5.2 Sac7d-E6AP for Yeast Display

DNA Sequence

5′-**GCTAGC**AGCGTGAAAGTGAAATTTCTGCTGAACGGCGAAGAAAAAGAAGTGGATA
CCAGCAAAATTCGCGATGTGAGTCGCCAGGGCAAAAACGTGAAATTTACCTATAACG
ATAACGGCAAATATGGCGCGGGCAACGTGGATGAAAAAGATGCGCCGAAAGAACTGC
TGGATATGCTGGCGCGC**GGATCC**-3′

Protein Sequence

Division between Sac7d R60 and E6AP E372 marked with //

N-**AS**SVKVKFLLNGEEKEVDTSKIRDVSRQGKNVKFTYNDNGKYGAGNVDEKDAPKELLD
MLAR//ELTKQELLGEER-C

C.5.3 E6AP Peptide for Yeast Display

DNA Sequence

5′-**GCTAGC**GAGCTGACTAAACAAGAACTTCTGGGCGAGGAGCGC**GGATCC**-3′

Protein Sequence

N-**AS**ELTKQELLGEER**GS**-C

C.5.4 sfGFP-E6

DNA Sequence

Beginning and end of ORF marked with //. NcoI and KpnI cloning sites in **bold**.

5′-**CCATG//G**GTTCTCATCACCATCATCACCACGCTAGCAAAGGTGAAGAGCTGTTTAC
GGGTGTAGTACCGATCTTAGTGGAATTAGACGGCGACGTGAACGGTCACAAATTTAG
CGTGCGCGGCGAAGGCGAAGGTGACGCTACCAATGGTAAATTGACCCTGAAGTTTAT
TTGCACAACAGGCAAATTACCCGTTCCGTGGCCCACCTTAGTGACCACCCTGACCTA
TGGCGTTCAGTGCTTCAGTCGTTACCCAGATCATATGAAACAACACGATTTTTTCAA
ATCAGCCATGCCTGAAGGATATGTTCAAGAGCGTACAATCAGCTTCAAGGACGATGG
CACCTATAAAACGCGTGCGGAAGTGAAATTTGAAGGCGACACATTAGTAAACCGTAT
CGAACTGAAAGGTATCGACTTCAAAGAAGACGGCAACATTTTAGGCCATAAGCTGGA
ATATAACTTTAATTCTCATAACGTGTATATTACGGCCGATAAACAGAAAAACGGTAT
CAAGGCAAATTTCAAAATTCGCCATAACGTGGAAGACGGCAGCGTTCAATTAGCGGA
TCATTATCAACAAAACACGCCGATTGGTGACGGGCCTGTACTGTTACCTGACAACCA
CTACCTGAGCACCCAGTCAGCACTGAGCAAAGATCCGAACGAAAAACGCGATCACAT
GGTTCTGTTAGAATTCGTGACCGCTGCAGGCATTACTCACGGAATGGACGAACTCTA
CAAGGCCGCAGCCTTTCAGGACCCACAGGAGCGACCCAGAAAGTTACCACAGTTATG
CACAGAGCTGCAAACAACTATACATGATATAATATTAGAATGTGTGTACTGCAAGCA
ACAGTTACTGCGACGTGAGGTATATGACCGTGCTTTTCGGGATTTATGCATAGTATA
TAGAGATGGGAATCCATATGCTGTATGTGATAAATGTTTAAAGTTTTATTCTAAAATT
AGTGAGTGTAGACATTATTCCTATAGTTTGTATGGAACAACATTAGAACAGCAATAC
AACAAACCGTTGGGTGATTTGTTAATTAGGTGTATTAACTGTCAAAAGCCACTGAGT
CCTGAAGAAAAGCAAAGACATCTGGACAAAAAGCAAAGATTCCATAATATAAGGGGT
CGGTGGACCGGTCGATGTATGTCTTGTAGCAGATCATCAAGAACACGTGGAGAACCC
AGCTGTAGAATTCCCTGCAGG//TAATAGGGTACC

Protein Sequence

sfGFP residue Ser2 in **bold**, E6 residue Phe2 also in **bold.** The end of sfGFP and the beginning of E6 sequences are marked with //

N-MGSHHHHHHA**S**KGEELFTGVVPILVELDGDVNGHKFSVRGEGEGDATNGKLTLKFICT
TGKLPVPWPTLVTTLTYGVQCFSRYPDHMKQHDFFKSAMPEGYVQERTISFKDDGTYKT
RAEVKFEGDTLVNRIELKGIDFKEDGNILGHKLEYNFNSHNVYITADKQKNGIKANFKIRH
NVEDGSVQLADHYQQNTPIGDGPVLLPDNHYLSTQSALSKDPNEKRDHMVLLEFVTAAGI
THGMDELYK//AAA//**F**QDPQERPRKLPQLCTELQTTIHDIILECVYCKQQLLRREVYDRAFR
DLCIVYRDGNPYAVCDKCLKFYSKISECRHYSYSLYGTTLEQQYNKPLGDLLIRCINCQKPLS
PEEKQRHLDKKQRFHNIRGRWTGRCMSCSRSSRTRGEPSCRIPCR-C

C.5.5 sfGFP

DNA Sequence

Beginning and end of ORF marked with //. NcoI and KpnI cloning sites in **bold**.

5′-**CCATG//G**GTTCTCATCACCATCATCACCACGCTAGCAAAGGTGAAGAGCTGTTTAC
GGGTGTAGTACCGATCTTAGTGGAATTAGACGGCGACGTGAACGGTCACAAATTTAG
CGTGCGCGGCGAAGGCGAAGGTGACGCTACCAATGGTAAATTGACCCTGAAGTTTAT
TTGCACAACAGGCAAATTACCCGTTCCGTGGCCCACCTTAGTGACCACCCTGACCTA
TGGCGTTCAGTGCTTCAGTCGTTACCCAGATCATATGAAACAACACGATTTTTTCAA
ATCAGCCATGCCTGAAGGATATGTTCAAGAGCGTACAATCAGCTTCAAGGACGATGG
CACCTATAAAACGCGTGCGGAAGTGAAATTTGAAGGCGACACATTAGTAAACCGTAT
CGAACTGAAAGGTATCGACTTCAAAGAAGACGGCAACATTTTAGGCCATAAGCTGGA
ATATAACTTTAATTCTCATAACGTGTATATTACGGCCGATAAACAGAAAAACGGTAT
CAAGGCAAATTTCAAAATTCGCCATAACGTGGAAGACGGCAGCGTTCAATTAGCGGA
TCATTATCAACAAAACACGCCGATTGGTGACGGGCCTGTACTGTTACCTGACAACCA
CTACCTGAGCACCCAGTCAGCACTGAGCAAAGATCCGAACGAAAAACGCGATCACAT
GGTTCTGTTAGAATTCGTGACCGCTGCAGGCATTACTCACGGAATGGACGAACTCTA
CAAGGCCGCAGCCTTTCAGGACCCACAGGAGCGACCCAGAAAGTTACCACAGTTATG
CACAGAGCTGCAAACAACTATACATGATATAATATTAGAATGTGTGTACTGCAAGCA
ACAGTTACTGCGACGTGAGGTATATGACCGTGCTTTTCGGGATTTATGCATAGTATA

TAGAGATGGGAATCCATATGCTGTATGTGATAAATGTTTAAAGTTTTATTCTAAAATT
AGTGAGTGTAGACATTATTCCTATAGTTTGTATGGAACAACATTAGAACAGCAATAC
AACAAACCGTTGGGTGATTTGTTAATTAGGTGTATTAACTGTCAAAAGCCACTGAGT
CCTGAAGAAAAGCAAAGACATCTGGACAAAAAGCAAAGATTCCATAATATAAGGGGT
CGGTGGACCGGTCGATGTATGTCTTGTAGCAGATCATCAAGAACACGTGGAGAACCC
AGCTGTAGAATTCCCTGCAGG//TAATAGGGTACC

Protein Sequence

sfGFP residue Ser2 in **bold**, E6 residue Phe2 also in **bold**. The end of sfGFP and the beginning of E6 sequences are marked with //

N-MGSHHHHHHA**S**KGEELFTGVVPILVELDGDVNGHKFSVRGEGEGDATNGKLTLKFICT
TGKLPVPWPTLVTTLTYGVQCFSRYPDHMKQHDFFKSAMPEGYVQERTISFKDDGTYKT
RAEVKFEGDTLVNRIELKGIDFKEDGNILGHKLEYNFNSHNVYITADKQKNGIKANFKIRH
NVEDGSVQLADHYQQNTPIGDGPVLLPDNHYLSTQSALSKDPNEKRDHMVLLEFVTAAGI
THGMDELYK//AAA//**F**QDPQERPRKLPQLCTELQTTIHDIILECVYCKQQLLRREVYDRAFR
DLCIVYRDGNPYAVCDKCLKFYSKISECRHYSYSLYGTTLEQQYNKPLGDLLIRCINCQKPLS
PEEKQRHLDKKQRFHNIRGRWTGRCMSCSRSSRTRGEPSCRIPCR-C

C.5.6 Sac7d for Expression

DNA Sequence

5′-ATGGGCAGCAGCCATCACCATCATCACCACAGCCAGGATCCCGTGAAAGTGAAAT
TTCTGCTGAACGGCGAAGAAAAAGAAGTGGATACCAGCAAAATTCGCGATGTGAGTC
GCCAGGGCAAAAACGTGAAATTTACCTATAACGATAACGGCAAATATGGCGCGGGCA
ACGTGGATGAAAAAGATGCGCCGAAAGAACTGCTGGATATGCTGGCGCGCGCGGAAC
GCGAAAAAAAACTGAACTTCTGGGCGAGGAGCGC||TAATAG**GGTACC**-3′

Protein Sequence

Bolded Val is canonically Val2 on Sac7d.

5′-MGSSHHHHHHSQDPVKVKFLLNGEEKEVDTSKIRDVSRQGKNVKFTYNDNGKYGAGN
VDEKDAPKELLDMLARAEREKKLN-3′

C.5.7 Sac7d-E6AP for Expression

DNA Sequence

ORF start and end designated with ||.

5′-||ATGGGCAGCAGCCATCACCATCATCACCACAGCCAGGATCCCGTGAAAGTGAAAT
TTCTGCTGAACGGCGAAGAAAAAGAAGTGGATACCAGCAAAATTCGCGATGTGAGTC
GCCAGGGCAAAAACGTGAAATTTACCTATAACGATAACGGCAAATATGGCGCGGGCA
ACGTGGATGAAAAAGATGCGCCGAAAGAACTGCTGGATATGCTGGCGCGCGAGCTGA
CTAAACAAGAACTTCTGGGCGAGGAGCGC||TAATAG**GGTACC**-3′

Protein Sequence

Bolded Val is canonically Val2 on Sac7d. Transition between Arg60 of Sac7d and Glu372 of E6AP designated with //.

5′-MGSSHHHHHHSQDP**V**KVKFLLNGEEKEVDTSKIRDVSRQGKNVKFTYNDNGKYGAGN
VDEKDAPKELLDMLAR//ELTKQELLGEER-3′

C.5.8 p53 Core

DNA Sequence

BamHI and KpnI cloning sites in **bold**. ORF start and end designated with ||. Primer overlap in *italics*.

5′-||ATGGGCAGCAGCCATCACCATCATCACCACAGCCA**GGATCC***C**AGCAGCAGCAGCGTGCC*
GAGCCAGAAAACCTATCAGGGCAGCTATGGCTTTCGCCTGGGCTTTCTGCATAGCGG
CACCGCGAAAAGCGTGACCTGCACCTATAGCCCGGCGCTGAACAAAATGTTTTGCCA
GCTGGCGAAAACCTGCCCGGTGCAGCTGTGGGTGGATAGCACCCCGCCGCCGGGCA
CCCGCGTGCGCGCGATGGCGATTTATAAACAGAGCCAGCATATGACCGAAGTGGTG

CGCCGCTGCCCGCATCATGAACGCTGCAGCGATAGCGATGGCCTGGCGCCGCCGCA
GCATCTGATTCGCGTGGAAGGCAACCTGCGCGTGGAATATCTGGATGATCGCAACA
CCTTTCGCCATAGCGTGGTGGTGCCGTATGAACCGCCGGAAGTGGGCAGCGATTGC
ACCACCATTCATTATAACTATATGTGCAACAGCAGCTGCATGGGCGGCATGAACCGC
CGCCCGATTCTGACCATTATTACCCTGGAAGATAGCAGCGGCAACCTGCTGGGCCGC
AACAGCTTTGAAGTGCGCGTGTGCGCGTGCCCGGGCCGCGATCGCCGCACC*GAAGAA*
GAAAACCTGCGCAAAAAA||TAATAG**GGTACC**-*3′*

Protein Sequence

The **bolded** Ser is canonically S94 in p53

N-MGSSHHHHHHSQDP**S**SSVPSQKTYQGSYGFRLGFLHSGTAKSVTCTYSPALNKMFCQL
AKTCPVQLWVDSTPPPGTRVRAMAIYKQSQHMTEVVRRCPHHERCSDSDGLAPPQHLIR
VEGNLRVEYLDDRNTFRHSVVVPYEPPEVGSDCTTIHYNYMCNSSCMGGMNRRPILTIITL
EDSSGNLLGRNSFEVRVCACPGRDRRTEEENLRKK-C

C.6 Sequences from Appendix B, "Other Experiments of Possible Interest"

C.6.1 ΔC25 TAR

5′-biotin-GCAGAUCGAGCCUGGGAGCUCUCUGC-3′

C.6.2 TEV-Cys-β2β3-Cys-TEV

DNA Sequence

NdeI and XhoI cloning sites in **bold**, the beginning of the in-frame TEV-Cys-β2β3-Cys-TEV sequence is designated with //

5′-**CATATG**GG//GAGAATCTATACTTCCAAAGCTGCGTGCCGCGCCAGCGCACCCCGCG
CGGCCAGGCGTGCGAAAACCTGTATTTTCAGAGCGCGT**CTCGAG**-3′

Protein Sequence

The G and A flanking this sequence are part of the NdeI and XhoI sites respectively.

TEV cleavage sites in *italics*

TBP 6.7 β2β3 loop in **bold**

N-G*ENLYFQS*C**VPRQRTPRGQAC***ENLYFQS*A-C

C.6.3 Cys-β2β3-Cys

DNA Sequence

NdeI and XhoI cloning sites in **bold**, the beginning of the in-frame Cys-β2β3-Cys sequence is designated with //

5′-**CATATG**GC//TGCGTGCCGCGCCAGCGCACCCCGCGCGGCCAGGCGTGCGCGT**CTC-GAG**-3′

Protein Sequence

The G and A flanking this sequence are part of the NdeI and XhoI sites respectively.

TBP 6.7 β2β3 loop in **bold**

N-GC**VPRQRTPRGQAC**A-C

C.6.4 Z-Peptide

DNA Sequence

NdeI and XhoI cloning sites in **bold**, the beginning of the in-frame Z-pep sequence is designated with //

5′-**CATATG**GC//CGCTGAAAAAAGAACTGCAGGCGAACAAAAAAGAACTGGCGCAGCTG AAATGGGAACTGCAGGCGCTGAAAAAAGAACTGGCGCAGG**CTCGAG**

Protein Sequence

The G and the A flanking this sequence are part of the NdeI and XhoI sites respectively.

Z-peptide sequence **bolded** N-G**ALKKELQANKKELAQLKWELQALKKELAQ**A-C

C.6.5 Gblock sequence from Section B.3, "TBP 6.7 Expresses in Mammalian Cells"

BsaI cloning sites **bolded**, associated overhangs *italicized*

5′-TATA**GGTCTC**TGATCATGGCACAGGTCCAACTGCAGGTTGACATGGCCGTCCCTGA
AACCCGGCCCAACCACACCATCTACATTAATAATCTGAACAGCAAGATTAAAAAGGA
CGAGCTCAAGAAATCTTTGTACGCTATTTTCTCACAGTTCGGTCAGATTTTGGATATC
CTCGTGCCTAGGCAGAGAACACCCCGGGGACAGGCCTTCGTAATATTTAAGGAGGTG
TCCTCAGCCACCAATGCCCTGCGGTCTATGCAAGGGTATCCATTTTACGACAAACCT
ATGAGGATTCAGTACGCTAGAACGGACAAAAGGATCCCTGCCAAGATGAAGGGCACC
TTCGGTAGCGTCGATAGCCGCGGCAGTCCAGTCGCCGCTGCCCTAGT**GAGACC**TATA
-3′

C.6.6 Primers from Section B.3, "TBP 6.7 Expresses in Mammalian Cells"

TBP 6.7 gBlock FP

5′-TATAGGTCTCTGATCATGGCACAGGTCCAACTGC-3′

TBP 6.7 * RP

5′-tataGGTCTCACTAGTCAGGCAGCGGCGACTGG-3′

TBP 6.7 FLAG RP

5′-AGGTCTCACTAGTCACTTATCGTCGTCATCCTTGTAATCGGCAGCGGCGACTGG-3′

TBP 6.7 NLS RP

5′-AGGTCTCTCTAGTTATCATTTCTTCTTTTTGGCTTGTCCTGCCTTTTTAGTGGCGGC
GGGGCGTTTGGCAGCGGCGACTGG-3′

C.6.7 TBP 6.7-*

DNA Sequence

BsaI overhangs *italicized*.

ORF designated with ||

5′-TATA**GGTCTC**T*GATC*||ATGGCACAGGTCCAACTGCAGGTTGACATGGCCGTCCCTGA
AACCCGGCCCAACCACACCATCTACATTAATAATCTGAACAGCAAGATTAAAAAGGA
CGAGCTCAAGAAATCTTTGTACGCTATTTTCTCACAGTTCGGTCAGATTTTGGATATC
CTCGTGCCTAGGCAGAGAACACCCCGGGGACAGGCCTTCGTAATATTTAAGGAGGTG
TCCTCAGCCACCAATGCCCTGCGGTCTATGCAAGGGTATCCATTTTACGACAAACCT
ATGAGGATTCAGTACGCTAGAACGGACAAAAGGATCCCTGCCAAGATGAAGGGCACC
TTCGGTAGCGTCGATAGCCGCGGCAGTCCAGTCGCCGCTGCC||TGA*CTAG*-3′

Protein Sequence

N-MAQVQLQVDMAVPETRPNHTIYINNLNSKIKKDELKKSLYAIFSQFGQILDILVPRQRTP
RGQAFVIFKEVSSATNALRSMQGYPFYDKPMRIQYARTDKRIPAKMKGTFGSVDSRGSPV
AAA-C

C.6.8 TBP 6.7-FLAG-*

DNA Sequence

BsaI overhangs *italicized*.

ORF designated with ||

5′-TATA**GGTCTC**T*GATC*||ATGGCACAGGTCCAACTGCAGGTTGACATGGCCGTCCCTGA
AACCCGGCCCAACCACACCATCTACATTAATAATCTGAACAGCAAGATTAAAAAGGA
CGAGCTCAAGAAATCTTTGTACGCTATTTTCTCACAGTTCGGTCAGATTTTGGATATC
CTCGTGCCTAGGCAGAGAACACCCCGGGGACAGGCCTTCGTAATATTTAAGGAGGTG
TCCTCAGCCACCAATGCCCTGCGGTCTATGCAAGGGTATCCATTTTACGACAAACCT
ATGAGGATTCAGTACGCTAGAACGGACAAAAGGATCCCTGCCAAGATGAAGGGCACC

TTCGGTAGCGTCGATAGCCGCGGCAGTCCAGTCGCCGCTGCCGATTACAAGGATGAC
GACGATAAG||TGA*CTAG-3′*

Protein Sequence

N-MAQVQLQVDMAVPETRPNHTIYINNLNSKIKKDELKKSLYAIFSQFGQILDILVPRQRTP
RGQAFVIFKEVSSATNALRSMQGYPFYDKPMRIQYARTDKRIPAKMKGTFGSVDSRGSPV
AAADYKDDDDK-C

C.6.9 TBP 6.7-NLS-*

DNA Sequence

BsaI overhangs *italicized*.

ORF designated with ||

5′-TATA**GGTCTC**T*GATC*||ATGGCACAGGTCCAACTGCAGGTTGACATGGCCGTCCCTGA
AACCCGGCCCAACCACACCATCTACATTAATAATCTGAACAGCAAGATTAAAAAGGA
CGAGCTCAAGAAATCTTTGTACGCTATTTTCTCACAGTTCGGTCAGATTTTGGATATC
CTCGTGCCTAGGCAGAGAACACCCCGGGGACAGGCCTTCGTAATATTTAAGGAGGTG
TCCTCAGCCACCAATGCCCTGCGGTCTATGCAAGGGTATCCATTTTACGACAAACCT
ATGAGGATTCAGTACGCTAGAACGGACAAAAGGATCCCTGCCAAGATGAAGGGCACC
TTCGGTAGCGTCGATAGCCGCGGCAGTCCAGTCGCCGCTGCCAAACGCCCCGCCGCC
ACTAAAAAGGCAGGACAAGCCAAAAAGAAGAAA||TGA*CTAG-3′*

Protein Sequence

Nuclear Localization Sequence **bolded**

N-MAQVQLQVDMAVPETRPNHTIYINNLNSKIKKDELKKSLYAIFSQFGQILDILVPRQRTP
RGQAFVIFKEVSSATNALRSMQGYPFYDKPMRIQYARTDKRIPAKMKGTFGSVDSRGSPV
AAA**KRPAATKKAGQAKKKK**-C

C.6.10 Sac7d Library

DNA Sequence

NheI, BamHI, and XhoI cloning sites **bolded**.

N-terminal *myc* tag *italicized*.

NNK codons corresponding to position 56, 59, 60, 63, and 64.

5′-**GCTAGC***GAACAAAAGCTTATTTCTGAAGAGGACTTG***GGATCC**GTGAAAGTGAAATTTCT
GCTGAACGGCGAAGAAAAAGAAGTGGATACCAGCAAAATTCGCGATGTGAGTCGCCA
GGGCAAAAACGTGAAATTTACCTATAACGATAACGGCAAATATGGCGCGGGCAACGT
GGATGAAAAAGATGCGCCGAAAGAACTGCTGNNKATGCTGNNKNNKGCGGAANNKN
NKAAAAAACTGAACTAATAG**CTCGAG**-3′

Protein Sequence

Canonical Sac7d Val2 **bolded**.

Randomized positions given by "X."

N-**V**KVKFLLNGEEKEVDTSKIRDVSRQGKNVKFTYNDNGKYGAGNVDEKDAPKELLXMLX
XAEXXKKLN-C

C.6.11 Sac7d with N-terminal *myc*

DNA sequence

NheI, BamHI, and XhoI cloning sites **bolded**.

N-terminal *myc* tag *italicized*

5′-**GCTAGC***GAACAAAAGCTTATTTCTGAAGAGGACTTG***GGATCC**GTGAAAGTGAAATTTCT
GCTGAACGGCGAAGAAAAAGAAGTGGATACCAGCAAAATTCGCGATGTGAGTCGCCA
GGGCAAAAACGTGAAATTTACCTATAACGATAACGGCAAATATGGCGCGGGCAACGT
GGATGAAAAAGATGCGCCGAAAGAACTGCTGGATATGCTGGCGCGCGCGGAACGCG
AAAAAAAACTGAACTGAACTAATAG**CTCGAG**-3′

Protein Sequence

Canonical Sac7d Val2 **bolded.**

N-**V**KVKFLLNGEEKEVDTSKIRDVSRQGKNVKFTYNDNGKYGAGNVDEKDAPKELLDML
ARAEREKKLN-C

C.6.12 CBP 5.21

DNA Sequence

NheI, BamHI, and XhoI cloning sites **bolded.**

N-terminal *myc* tag *italicized.*

Bases that are the result of library screen in ***bolded italics.***

5′-**GCTAGC***GAACAAAAGCTTATTTCTGAAGAGGACTTG***GGATCC**GTGAAAGTGAAATTTCT
GCTGAACGGCGAAGAAAAAGAAGTGGATACCAGCAAAATTCGCGATGTGAGTCGCCA
GGGCAAAAACGTGAAATTTACCTATAACGATAACGGCAAATATGGCGCGGGCAACGT
GGATGAAAAAGATGCGCCGAAAGAACTGCTG***TCG***ATGCTG***CGTACG***GCGGAAACTTTT
AAAAAACTGAACTAATAG**CTCGAG**-3′

Protein Sequence

Canonical Sac7d Val2 at start of sequence and all results of library screening (positions 56, 59, 60, 63, and 64) **bolded.**

N-**V**KVKFLLNGEEKEVDTSKIRDVSRQGKNVKFTYNDNGKYGAGNVDEKDAPKELL**S**ML
RTAET**F**KKLN-C