Title
Arginine-rich motif peptides as tools for understanding single-stranded DNA recognition

Author
Landt, Stephen George

Publication Date
2004

Peer reviewed|Thesis/dissertation

Arginine-rich motif peptides as tools for understanding single-stranded DNA recognition

by

Stephen George Landt

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Biochemistry and Molecular Biology

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

[redacted]

I dedicate this thesis to Walter Ellett and Louis Landt,

two people I hoped would be here when I completed this work

Acknowledgements

This section has to begin with a heartfelt thank you to my wonderful wife Denise. She has been with me longer than graduate school has and she has been supportive the whole time, even though it has probably kept us from doing some things she wanted to do. I also want to thank my parents for their support throughout the whole process and for taking pride in my accomplishments. That means a lot to me. I would also like to thank the family I have married into. Denise's mom and grandma have been loving and supportive every day that I've known them.

Of course, I owe a great deal to my advisor, Alan Frankel. In talking to other students and watching other groups, I've come to realize that Alan has a genuine concern for the people in his lab that is uncommon. As with any scientific relationship that has lasted as long as ours has, there have been moments when I've questioned the direction in which things were going, but, looking back, I think Alan's willingness to support anything that I or anyone else has brought enthusiasm to has made my time in the lab enjoyable and has helped me identify the type of scientist I would like to become.

I've had a chance to work with a whole lot of great people in the lab. So many, in fact, that I am wary of trying to list them all and instead will offer them a great big collective thanks. However, I would like to thank a couple of people personally. Rob Nakamura has a been a great person to work with and to have around as a friend. Although the project may have beaten us up a bit, it was much more enjoyable because of the collaboration. The other person I would like to thank is Colin Smith, who was invaluable to me during my early years in grad school. Looking back, I think much of

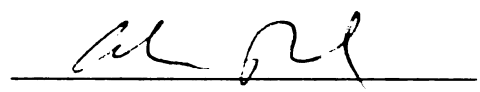
how I do things and interact with people scientifically today stems from trying to imitate the way Colin dealt with me.

Finally, I'd like to thank the members of my thesis committee- Liz Blackburn, Raul Andino, and Peter Walter. I've enjoyed your interest in my work and your help in guiding me, especially during these last couple of years.

Arginine-rich motif peptides as tools for understanding single-stranded DNA recognition

by

Stephen G. Landt



Alan D. Frankel, Ph.D.

Thesis advisor and thesis committee chair

Although DNA is generally found in the double-stranded form, many of the important events that happen to it occur while it is single-stranded. In many of these cases, it is expected that the recognition of specific single-stranded DNA (ssDNA) sequences by proteins will be essential for the fulfillment of ssDNA function. Very little is known about the mechanisms of ssDNA recognition by proteins, but much of what is known has come from structural studies on the specific binding of linear, unstructured DNA sequences by what are conventionally thought of as "RNA-binding domains". Given the requirement that any ssDNA sequence in the cell must compete with the duplex form for its existence, it is likely that stable secondary and tertiary structures will also be important for ssDNA recognition. By analogy to linear ssDNA recognition, the study of protein domains specific for structured RNA should prove valuable for understanding how structured ssDNA is recognized.

This thesis describes the use of the arginine-rich motif (ARM) family of RNA binding proteins to understand ssDNA recognition. Initially, we show that, while the affinity of

at least one ARM:RNA interaction depends strongly on features of the A-form RNA helix, this dependence is localized and that multiple energetically critical protein:RNA contacts appear to be viable in the B-form helix geometry that is characteristic of DNA. We expand on this by using in vitro selection to identify high affinity ssDNA ligands to the ARM peptide from the HIV Rev protein. We show that Rev can bind ssDNA with affinities and specificities that compare favorably to its ability to recognize RNA. We also characterize a mechanism for the recognition of branched nucleic acids in which an aromatic amino acid sidechain appears to stack on the end of one DNA helix in the context of a 3-helix junction structure that emerged from our selection. Finally, we identify a simple interaction motif, a 5' G•T/CG basestep that is recognized by two arginine separated by one turn of an α-helix, which should facilitate specific ssDNA binding by a variety of nucleic acid binding proteins to ssDNA in a range of structural contexts.

Table of Contents

Chapter 1

General Introduction

Most cellular processes that involve nucleic acids are controlled by the binding of specific proteins. For DNA, this is true for the execution and regulation of replication, transcription, and many aspects of recombination and repair. It is also often true for RNA, from the regulation of splicing and translation to the control of subcellular localization and degradation and, in the case of RNA viruses, replication. Because of this, much effort has gone towards identifying the proteins that specifically bind DNA and RNA in these processes and in studying how they recognize their targets amid the sea of other nucleic acids in the cell. This has resulted in the identification of an array of different protein families, generally classified by their sequence and structural similarities to one another.

An additional classification that has taken hold from these studies is the distinction between DNA-binding proteins and RNA-binding proteins. This assignment is usually based on the circumstances under which the founding or most prominent member of a protein family was characterized and is often useful in the understanding the biological functions of the family. However, this simple categorization overlooks a degree of overlap between DNA-binding and RNA-binding activities within a family that is generally underestimated. In fact, for some of the most abundant types of RNA-binding domains known, family members have been identified that bind DNA or RNA. For instance, while the RNA-recognition motif (RRM) is generally considered an RNA-binding domain, hnRNP A1 has been shown to specifically associate with telomeric DNA through its two RRM domains [1]. HnRNP K, the titular head of the K-homology (KH) domain family of RNA-binding proteins, has also now been shown to function as a transcriptional activator through its high affinity interaction with a target DNA site [2]. In

addition, there are many examples of well-known proteins that have been shown to bind both RNA and DNA with high-affinity, with biological functions proposed for both activities [3].

In these families, for cases where nucleic acid recognition is understood at the structural level, it is generally observed that DNA and RNA are bound by the same surfaces of the molecules, which essentially invalidates the DNA-binding vs. RNA-binding distinction for these groups. However, comparison of the nucleic acid conformations in these complexes shows that the amount of secondary or tertiary structure is generally similar in the different binding sites. KH-domain RNA substrates are generally unfolded linear sequences, such as the branchpoint sequence recognized by the BBP/SF1 protein in the control of mRNA splicing, as are the single-stranded poly-pyrimidine stretches that are the targets of hnRNP K in the c-myc promoter DNA [1,4-6]. The same is true for RRM proteins: U1A, the prototype RRM protein, recognizes the 10 nucleotide loop of an RNA hairpin in an extended conformation, where at least three of the positions are important solely because they allow conformational flexibility in the binding site and the crystal structure of hnRNP A1 in complex with a DNA oligonucleotide containing a telomeric repeat shows that the DNA is also bound in an extended conformation [1,7].

This suggests that it may be useful to think of nucleic acid binding sites as a structural continuum ranging from the fully linear, lacking any significant intramolecular interactions, to the fully basepaired. Protein families can be classified by the location of their binding sites along this range. The KH and RRM families, for instance, would be classified as binders of nucleic acids with little secondary or tertiary structure. These

classifications also reflect on the mechanism of recognition. Molecules which recognize unstructured nucleic acid have been shown to use stacking of hydrophobic residues with nucleic acid bases and hydrogen bonds to all faces of the bases and backbone to achieve specificity, while proteins which recognize double-stranded nucleic acids do not have access to these strategies and instead recognize the non Watson-Crick faces of nucleotide bases as well as backbone features associated with different helix geometries [8,9].

While this logic is quite useful for less structured nucleic acids, it is unclear how valuable it is for nucleic acids with extensive secondary and tertiary structure. This is primarily because of the fundamental differences between the parent helical forms of DNA and RNA. The B-form DNA helix has a much shallower, wider major groove than does the A-form RNA helix, which allows easy access to units of protein structure that recognize the pattern of chemical groups unique to the bases in a given sequence. Additionally, the 2'OH group of RNA is accessible to proteins in the minor groove, and the change in backbone sugar conformation that arises as a consequence of the 2'OH creates a very different distribution of hydrophobic surfaces along the helix backbones [10]. These are substantial differences, which might be expected to require the evolution of protein binders without the cross-reactivity between DNA and RNA that is seen for unstructured molecules.

Indirect evidence, however, does suggest that protein families known to bind B-form DNA can, in some cases, profit from deviations in their binding sites towards the A-form geometry. For example, it has been determined from crystal structures that a variety of DNA binding proteins, belonging to many different domain families, recognize target sites with increased major groove depths and basepair inclinations that are more like A-

form DNA than B-form DNA [11]. There is also the case of the ZF-QQR zinc-finger protein, which recognizes its target sequence ~15-fold better when it is presented in the context of an RNA-DNA hybrid, and presumably in a conformation intermediate between the A- and B- forms, than as double-stranded DNA (dsDNA) [12]. These limited examples raise the questions of quantitatively how specific double-stranded nucleic acid binding proteins are for a particular helix geometry and what the molecular mechanisms underlying this specificity are. Do proteins adapt cooperatively to many elements of a helix conformation across the entire protein:nucleic acid interface, or is the apparent specificity the result of a small number of specific contacts to discrete features of a given helix geometry?

The key to addressing these questions lies in locally altering the geometry of a target nucleic acid helix to see if certain regions are more important in determining protein specificity. In chapter 2, I describe the use of selective replacement of backbone deoxyriboses with riboses to locally modulate the helical conformation of a DNA version of the Bovine Immunodeficiency Virus (BIV) TAR RNA, the recognition site for the BIV Tat protein[13,14]. BIV Tat binding is determined by a 17 amino acid peptide that is classified with the arginine-rich motif (ARM) family of proteins, a group of short, highly basic peptides that bind RNA specifically and are thought to require the deep A-form major groove to achieve a binding interface large enough for high affinity[15]. We have also used the same method to analyze the interaction between the free amino acid arginine and the TAR RNA from HIV. This interaction is of much lower affinity, but appears to mimic several of the key features responsible for specificity in the BTat:BTAR complex [16-21]. Our results indicate that, while elements of the BTat:BTAR complex

that are in common with the arginine:HTAR interaction are very specific for the A-form, some of the most important contacts occur to regions that do not show a preference for one geometry. This mosaicity suggests that many of the strategies used to recognize RNA may also be useful for recognizing DNA and that the boundaries between other dsDNA-binding motifs and structured RNA-binding motifs may also be fuzzy.

The idea of a continuum of nucleic acid recognition is useful in one additional regard. It illustrates that there are very few examples of proteins that recognize DNA that is highly structured but not perfectly double-helical. In large part this is a consequence of the biology of DNA. Since it always exists in the presence of its complementary strand, except in specific cases such as at telomeres or in single-stranded DNA (ssDNA) viruses, the perfectly basepaired form will usually be available as a kinetically stable, energetically favorable alternative to almost any imperfectly paired structure.

However, I would argue that these sorts of structures may sometimes be more biologically accessible than is currently appreciated. DNA in the cell is under a variety of stresses that can alter the kinetics and thermodynamics of alternate structure formation. These factors include the inherent negative supercoiling of chromosomal DNA, as well as the further strain introduced by the unwinding that accompanies replication and transcription [22-24]. In theory, these forces can all catalyze the formation of alternate DNA structures and in practice, there are well established examples of biological functions that depend on ssDNA extrusion from the chromosome. These include transcriptional activation mediated by hnRNP K and its family members, as well as several other examples of ssDNA-based transcriptional activators and the proposed role for cruciform binding proteins in controlling replication initiation[25-29]. In addition, the

recognition of ssDNA structures could, at least in theory, be central to many elegant regulatory circuits. For instance, it is known that in prokaryotes, the level of supercoiling in the cell increases under environmental stresses, such as osmotic shock or the presence of antibiotics, suggesting that increased levels of ssDNA could act as a sensor within the chromosome of stress levels that could be rapidly acted upon by specific binding proteins to activate transcription [23]. Since both transcription and replication can promote ssDNA in regions immediately adjacent to the locations of these activities, it is possible that the recognition of ssDNA induced by these processes could serve either as an amplification mechanism for activating genes that are basally transcribed or for coupling transcription temporally to DNA replication[24]. Based on these considerations, it is more likely that ssDNA structures in the cell are poorly understood because they exist only for short times and often under special conditions, making them difficult to identify and study, rather than because they do not exist at all.

With this perspective, it is worthwhile to consider the advantages that structured ssDNA sites offer for protein recognition. A general lesson that can be drawn from studying RNA recognition is that the tightest binding sites tend to have at least some amount of secondary or tertiary structure. In some cases, this is minimal. For instance, in the very high affinity interaction between the U1A protein and its hairpin binding site on the U1 snRNA, only a single basepair is needed for recognition, but it provides enough constraints on the bases in the 10-nucleotide loop to be important for tight binding [6]. Secondary structure can also be substantial, as seen for ARM-family recognition. In this case, binding sites are always extensively basepaired and interrupted helices are stacked, illustrated at in the extreme by the ability of BIV Tat to bind tightly

to an RNA double-helix containing just a single base-bulge[15,30]. Much more elaborate tertiary structures are also common, structures involving the junctions of three or four helices and the juxtaposition of regions that are distant in primary structure, as exemplified by many of the protein interactions with ribosomal RNA [31,32]. In all these cases, the nucleic acid sites balance the primary advantage of single-stranded nucleic acid recognition- tremendous conformational diversity, including unique arrangements that can be very specifically recognized- with a primary advantage of double-stranded sites- a structural rigidity that minimizes the conformational entropy lost upon binding. This combination, which is accessible to RNA but is not well characterized in DNA structures, explains why the measured binding affinities of RNA binding domains are usually higher than are those of dsDNA binding domains and why these latter proteins generally function as dimers or as arrays of multiple repeats of a single motif.

The later chapters of this thesis describe my efforts to create, by in vitro selection, model systems for protein recognition of ssDNA and to biochemically characterize the mechanisms of recognition in these systems. Most of this work has been carried out with the ARM peptide from the HIV Rev protein as the protein target, in order to bias my selection for ligands that use extensive secondary and/or tertiary structure to recognize their target. I am particularly interested in identifying DNA sequence and structure motifs which are especially capable of facilitating tight binding to proteins and in characterizing the interplay between structure specificity and sequence specificity in these models. Intriguingly, we have selected two different classes of ssDNA molecules, one a small hairpin that is extensively basepaired, but with few elements of tertiary structure and another which adopts a complex 3-helix junction fold. These results,

however, also identify a common motif, the basestep consisting of the non Watson-Crick G•T basepair atop a CG basepair, as a unifying feature, which may be generally important for the recognition of ssDNA.

References

1. Ding, J., Hayashi, M., Zhang, Y., Manche, L., Krainer, AR & Xu, R. Crystal structure of the two-RRM domain of hnRNP A1 (UP1) complexed with single-stranded telomeric DNA. *Genes and Development* **13**, 1102-1115 (1999).

2. Michelotti, E., Michelotti, G., Aronsohn, A. & Levens, D. Heterogeneous nuclear ribonucleoprotein K is a transcription factor. *Molecular and Cellular Biology* **16**, 2350-2360 (1996).

3. Cassiday, L. & Maher, L. Having it both ways: transcription factors that bind DNA and RNA. *Nucleic Acids Research* **30**, 4118-4126 (2002).

4. Braddock, D., Louis, J., Baber, J., Levens, D. & Clore, G. Structure and dynamics of KH domains from FBP bound to single-stranded DNA. *Nature* **415**, 1051-1056 (2002).

5. Lewis, H.A. *et al.* Sequence-specific RNA binding by a Nova KH domain: implications for paraneoplastic disease and the Fragile X syndrome. *Cell* **100**, 323-332 (2000).

6. Oubridge, C., Ito, N., Evans, P.R., Teo, C.H. & Nagai, K. Crystal structure at 1.92 A resolution of the RNA-binding domain of the U1A spliceosomal protein complexed with an RNA hairpin. *Nature* **372**, 432-438 (1994).

7. Williams, D.J. & Hall, K.B. RNA hairpins with non-nucleotide spacers bind efficiently to the human U1A protein. *J Mol Biol* **257**, 265-275 (1996).

8. Nagai, K. RNA-protein complexes. *Curr Opin Struct Biol* **6**, 53-61 (1996).

9. Theobald, D., Mitton-Fry, R. & Wuttke, D. Nucleic acid recognition by OB-fold proteins. *Annual Review of Biophysics and Biomolecular Structure* **12**, 794-801 (2003).

10. Saenger, W. *Principles of nucleic acid structure*, (Springer-Verlag, 1984).

11. Nekludova, L. & Pabo, C.O. Distinctive DNA conformation with enlarged major groove is found in Zn-finger-DNA and other protein-DNA complexes. *Proc Natl Acad Sci U S A* **91**, 6948-6952 (1994).

12. Shi, Y. & Berg, J.M. Specific DNA-RNA hybrid binding by zinc finger proteins. *Science* **268**, 282-284 (1995).

13. Chen, L. & Frankel, A.D. An RNA-binding peptide from bovine immunodeficiency virus Tat protein recognizes an unusual RNA structure. *Biochemistry* **33**, 2708-2715 (1994).

14. Chen, L. & Frankel, A.D. A peptide interaction in the major groove of RNA resembles protein interactions in the minor groove of DNA. *Proc. Natl. Acad. Sci. USA* **92**, 5077-5081 (1995).

15. Weiss, M. & Narayana, N. RNA recognition by arginine-rich peptide motifs. *Biopolymers* **48**, 167-180 (1999).

16. Aboul-ela, F., Karn, J. & 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**, 313-332 (1995).

17. Brodsky, A.S. & Williamson, J.R. Solution structure of the HIV-2 TAR-argininamide complex. *J. Mol. Biol.* **267**, 624-639 (1997).

18. Long, K.S. & Crothers, D.M. Characterization of the solution conformations of unbound and Tat peptide-bound forms of HIV-1 TAR RNA. *Biochemistry* **38**, 10059-10069 (1999).

19. Puglisi, J.D., Tan, R., Calnan, B.J., Frankel, A.D. & Williamson, J.R. Conformation of the TAR RNA-arginine complex by NMR spectroscopy. *Science* **257**, 76-80 (1992).

20. Puglisi, J.D., Chen, L., Blanchard, S. & Frankel, A.D. Solution structure of a bovine immunodeficiency virus Tat-TAR peptide-RNA complex. *Science* **270**, 1200-1203 (1995).

21. Ye, X., Kumar, R.A. & Patel, D.J. Molecular recognition in the bovine immunodeficiency virus Tat peptide-TAR RNA complex. *Chem. Biol.* **2**, 827-840 (1995).

22. McClellan, J., Boublikova, P., Palecek, E. & Lilley, D. Superhelical torsion in cellular DNA responds directly to environmental and genetic factors. *Proceedings of the National Academy of Sciences* **87**, 8373-8377 (1990).

23. Dayn, A. *et al.* Formation of (dA•dT)n cruciforms in Escherichia coli cells under different environmental conditions. *Journal of Bacteriology* **173**, 2658-2664 (1991).

24. Dayn, A., Malkhosyan, S. & Mirkin, S. Transcriptionally driven cruciform formation in vivo. *Nucleic Acids Research* **20**, 5991-5997 (1992).

25. Dai, X., Greizerstein, M., Nadas-Chinni, K. & Rothman-Denes, L. Supercoil-induced extrusion of a regulatory DNA hairpin. *Proceedings of the National Academy of Sciences* **94**, 2174-2179 (1997).

26. Novac, O., Alvarez, D., Pearson, C., Price, G. & Zannis-Hadjopoulos, M. The human cruciform-binding protein, CBP, is involved in DNA replication and associates in vivo with mammalian replication origins. *Journal of Biological Chemistry* **277**, 11174-11183 (2002).

27. Spiro, C., Bazett-Jones, D., Wu, X. & McMurray, C. DNA structure determines protein binding and transcriptional efficiency of the proenkephalin cAMP-responsive enhancer. *Journal of Biological Chemistry* **270**, 27702-27710 (1995).

28. Tomonaga, T. & Levens, D. Activating transcription from single stranded DNA. *Proceedings of the National Academy of Sciences* **93**, 5830-5835 (1996).

29. Tomonaga, T. & Levens, D. Heterogeneous Nuclear Ribonucleoprotein K is a DNA-binding transactivator. *Journal of Biological Chemistry* **270**, 4875-4881 (1995).

30. Smith, C.A., Crotty, S., Harada, Y. & Frankel, A.D. Altering the context of an RNA bulge switches the binding specificities of two viral Tat proteins. *Biochemistry* **37**, 10808-10814 (1998).

31. Nikulin, A. *et al.* Crystal structure of the S15-rRNA complex. *Nature Structural Biology* **7**, 273-277 (2000).

32. Agalarov, S., Prasad, G., Funke, P., Stout, C. & Williamson, J. Structure of the S15,S6,S18-rRNA complex: assembly of the 30S ribosome central domain. *Science* **288**, 107-112 (2000).

Chapter 2

Localized influence of 2' hydroxyl groups and helix geometry on protein recognition in the RNA major groove

Localized influence of 2' hydroxyl groups and helix geometry on protein recognition in the RNA major groove

Stephen G. Landt, Alicia R. Tipton, and Alan D. Frankel*

[1]Department of Biochemistry and Biophysics
University of California, San Francisco
San Francisco, CA 94143-2280

Running Title: 2' OH groups and RNA major groove recognition

*Address correspondence to: Alan Frankel
Department of Biochemistry and Biophysics
UCSF
600 16th Street
San Francisco, CA 94143-2280

Telephone: 415-476-9994
FAX: 415-514-4112
e-mail: frankel@cgl.ucsf.edu

Abstract

The influence of the 2' hydroxyl group on the ribose sugar conformation and the overall RNA double helix geometry are well known but the effects on protein recognition are less well understood. To begin to examine how 2' hydroxyl groups might affect local or global helix structure and subsequent protein binding, we generated a series of DNA analogs of HIV and BIV TAR RNAs in which ribose sugars were systematically substituted in and around the known binding sites for argininamide and a BIV Tat arginine-rich peptide, respectively, and measured their corresponding binding affinities. For each TAR interaction, binding occurs in the RNA major groove with high specificity whereas binding to the all-DNA analog is weak and nonspecific. Relatively few substitutions are needed to convert either DNA analog of TAR into a high-affinity binder, with the ribose requirements being restricted largely to regions that directly contact the respective ligand. Substitutions at individual positions show up to 70-fold differences in binding affinity, even at adjacent base pairs, while two base pairs at the core of the BIV peptide-RNA interface are largely unaffected by deoxyribose substitution. The results suggest that the helix geometries and unique conformational features required for binding are established locally and are relatively insulated from effects more than one base pair away. It seems plausible that arginine-rich peptides are able to adapt to a mosaic helical architecture in which segments as small as single base steps may be considered as modular recognition units.

Introduction

Sequence-specific recognition of nucleic acids is commonly achieved by the complementary positioning of hydrogen bonding groups of the protein side chains and backbone with those of the bases, base pairs, or base pair steps in the major groove of a double helix [1,2]. Thus, the form of the double helix (A, B, or other) can influence recognition by altering the relative positioning of donors and acceptors, as can the presence of local helical irregularities, bends, or nonhelical structures. Indeed, an analysis of DNA-protein cocrystal structures has shown that many protein binding sites diverge significantly from the canonical B-form geometry, sometimes towards the A-form and often in ways that increase complementarity between the major groove and a protein α-helix[3,4]. The importance of recognizing sequence in an "A-like" geometry was further suggested by the observation that a zinc finger protein binds more tightly to its target sequence when presented in the context of an RNA-DNA hybrid than DNA[5]. While some proteins may recognize sequences within A-form conformations, an extended A-form duplex presents a deep and narrow major groove that is too small to accommodate an α-helix or β-sheet portion of a protein[6]. Thus, recognition of base pairs within an RNA major groove typically is accompanied by local interruptions of helical structure, having adjacent bulged or looped nucleotides that widen the groove sufficiently for protein access [7,8].

One type of RNA-binding domain that recognizes the RNA major groove is the arginine-rich motif (ARM), a 10-20 amino acid segment enriched in arginine that typically binds RNA with high affinity and specificity as an isolated peptide[9]. ARM

peptides can adopt α-helical, β-hairpin, or extended conformations and often require RNA to assume their bound conformations. It is likely that these peptides must conform to the structure of the deep major groove in order to create a sufficient interface for specific, high affinity binding[10,11]. Among the most well-characterized ARM peptide-RNA interactions are the Tat-TAR complexes from the human and bovine immunodeficiency viruses (HIV and BIV). The viral Tat proteins are essential transcription elongation factors that bind to their respective TAR RNA hairpins located at the 5' ends of the nascent transcripts[12].

In the HIV Tat-TAR complex, an ARM peptide comprised of amino acids 49-57 binds TAR with high affinity and primarily uses one arginine (at position 52) to make specific contacts to the RNA[7,12-15]. The free amino acid arginine, or even the guanidinium group alone, also binds TAR specifically, and NMR structures of TAR-argininamide complexes have shown that arginine and Tat peptides induce similar conformational changes in the RNA in which the helices surrounding a 3-nucleotide bulge coaxially stack and the bulged U23 residue is positioned in the major groove of the upper stem to form a base triple with the A27•U38 pair (Fig.1A)[16-21]. Arginine, which appears to be presented from an extended peptide chain, stacks between A22 and U23, hydrogen bonding to the O6 and N7 atoms of G26 in the major groove, and forms electrostatic or hydrogen bonding contacts with the phosphates of A22 and U23. This arginine-binding motif has shown up repeatedly in selection experiments with arginine or ARM peptides, suggesting that it represents a highly favorable major groove interaction[22-25].

Despite significant similarities between the HIV and BIV hairpin sequences, including conservation of the arginine-binding site, the BIV Tat-TAR interaction is far more extensive (Fig. 1A)[26-28]. The BIV interaction involves 8 critical peptide residues and a more extended RNA binding region[26,27]. The peptide is unstructured on its own but adopts a β-hairpin upon binding to the major groove, as shown by NMR structures of the complex [29,30]. A base triple similar to that in HIV TAR is observed, with arginine 73 making analogous contacts, however the A13•U24 base pair of the triple can be mutated to U•A with little effect on binding affinity[27]. A critical isoleucine at position 79 appears to buttress the U10 bulged base by stacking against a hydrophobic C5-C6 surface, and additional contacts between arginine 70 and G14, glycine 71 and G22, threonine 72 and the G22/C23 step, and potentially between arginine 77 and G9 at the top of the lower stem all contribute to the high affinity interaction.

For DNA-binding proteins there is substantial evidence that helical conformation plays an important role in recognition but little is known about how the form of the helix may contribute to RNA binding[3]. To examine the roles of helix geometry in ARM peptide recognition, we systematically introduced ribose groups into DNA analogs of HIV and BIV TAR, which do not bind argininamide or the BIV Tat peptide, respectively, and identified positions critical for binding. Because the 2' hydroxyl groups are the major determinant of A-form helix geometry and project into the minor groove where they inaccessible to ARM peptides bound in the major groove, we reasoned that such DNA/RNA hybrids might reveal which portions of the binding sites are particularly sensitive to helical conformation[6]. We show that the ribose requirements are highly localized to particular regions of the binding sites whereas other immediately adjacent

portions of the sites are largely insensitive. Thus, some essential peptide-RNA contacts appear to require the A-form geometry whereas others may adapt to slightly different helical forms, consistent with the conformational plasticity of ARM peptides.

Materials and Methods

Modeling

Canonical A- and B-form helices with the sequences of HIV and BIV TAR were generated by the Nucleic Acid Builder software package [31]. Using InsightII, these helices were superimposed on the average NMR structures of HIV TAR:argininamide[18] and BIV Tat:BIV TAR[29], (PDB ID 1MNB) complexes by overlapping all nitrogen atoms of the four bases at the junction between the upper and lower TAR stems (the A22:U40 and G26:C39 base pairs for HIV TAR and the G9:C26 and G11:C25 base pairs for BIV TAR).

Preparation and characterization of RNA, DNA and DNA/RNA hybrids

RNA was prepared by in vitro transcription, purified, and labeled as described [27]. DNA and DNA/RNA hybrids were synthesized on an Applied Biosystems Model 394 oligonucleotide synthesizer using phosphoramidites from Glen Research (Sterling, VA) and cleaved and deprotected according to manufacturer's protocols. Briefly, hybrids were deprotected overnight in ethanolic ammonium hydroxide and evaporated to dryness. Following 24 hr incubation in triethylamine trihydrofluoride to remove the silyl protecting groups, oligonucleotides were precipitated once with 0.3M sodium acetate, pH 6.0 and 3 volumes 1-butanol, and once with ethanol. Hybrids were purified on 15% polyacrylamide/urea gels and stored at –80° in deionized water. DNA and DNA/RNA hybrids were end-labeled with ^{32}P as described [27] and the positions of ribonucleotide

substitutions were confirmed by hydrolysis in 33mM sodium bicarbonate, pH 9.0 for 15 min at 90°C followed by analysis on 20% polyacrylamide/urea gels.

L-arginine-affinity chromatography

For HIV TAR analog binding experiments, L-arginine agarose columns (1 mL of 8.4 μmol/mL packed resin in 10 mL disposable columns; Sigma Chemical Co.) were pre-washed with 5 mL of 1 M NaCl/1X binding buffer (10 mM Tris·HCl pH 7.2, 0.2 mM EDTA), followed by 5 mL of 100 mM NaCl/1X binding buffer. A mixture of ^{32}P-5' end labeled RNA or hybrid molecule (~300,000 cpm) and 10μg of TAR DNA was loaded onto the column at 4°C in 100 mM NaCl/1X binding buffer and oligonucleotides were eluted with a 100-500 mM NaCl gradient (100 ml) at 4°C. Fractions (1.2 mL) were collected and analyzed on 15% polyacrylamide/urea gels to determine elution volumes. Dissociation constants (K_ds) were estimated using a standard curve relating K_ds measured by isocratic elution to the concentration of NaCl required for elution [Tao, 1996, 1997]. For calibration, HIV TAR RNA elutes at 165 mM NaCl whereas the DNA analog elutes at 79 mM NaCl.

BIV Tat peptide affinity chromatography

An 18-amino acid peptide corresponding to residues 65-81 of BIV Tat following a cysteine (Cys-BTat; CSGPRPRGTRGKGRRIRR) was synthesized and purified as described [26] and quantified by reactivity of the thiol group to Ellman's reagent [32]. The BTat affinity resin was generated by first activating 1 mL of packed ω-amino-hexyl agarose (4% agarose, epoxy activated with 12 atom spacer; Sigma) with freshly prepared

2.5 mM Sulfo-SMCC (Pierce) in 5 mL 50mM sodium phosphate, pH 7.4 for 30 min at room temperature. Activated agarose was washed 3 times with 30 mL of 50 mM sodium phosphate, pH 7.4 and resuspended in 2 mL 50 mM sodium phosphate, pH 7.4 containing 50 μg Cys-BTat. The peptide:agarose mixture was incubated for 2 hr at room temperature and unreacted resin was blocked by incubating with 20 μL of 500 mM dithiothreitol for an additional 30 min. Resin was washed 3 times with 50 mM sodium phosphate, pH 7.4 and resuspended in 1mL.

For BIV TAR analog binding experiments, a BTat column (333 μL packed resin in a 10 mL disposable column) was washed with 5 column volumes of 900 mM NaCl/1X binding buffer (10 mM Tris-Cl pH 7.4, 1 mM EDTA, 0.005% Triton X-100) followed by 5 column volumes of 100 mM NaCl/1X binding buffer. ^{32}P-5' end labeled RNA or hybrid (>100,000 cpm) was heated to 85°C in 1 mL of 100 mM NaCl/1X binding buffer and slow cooled to 4°C, 50 μg yeast tRNA (Invitrogen) was added, the mixture was loaded onto the column, and oligonucleotides were eluted with a 100-900 mM NaCl/1X binding buffer gradient (80mL) at 4°C. Elution volumes were determined by scintillation counting, and results from at least three independent experiments were averaged. K_ds were determined using a standard curve relating K_ds measured by fluorescence anisotropy (see below) to the concentration of NaCl required for elution. A tight linear relationship was obtained for all hybrids with measurable K_ds from anisotropy experiments.

Fluorescence anisotropy

Cys-BTat was labeled with fluorescein at its N-terminus by incubating 50 μM peptide with 500 μM 5-(iodoacetamido)fluorescein in 20 mM sodium phosphate pH, 8.0, 2 mM

EDTA for 2 h at room temperature in the dark. Labeled peptide was purified by C_4 reverse-phase HPLC as described for the unlabeled peptide [26] and was quantified by fluorescein absorbance at 475 nm. To measure binding, fluorescein-labeled Cys-BTat (2.5 nM) was incubated with varying concentrations of DNA, RNA, or DNA/RNA hybrids (0.5 – 8092 nM) in binding buffer (30 mM Hepes pH 7.5, 100 mM KCl, 40 mM NaCl, 10 mM ammonium acetate, 10 mM guandinium, 2 mM $MgCl_2$, 0.5 mM EDTA, 0.001% Nonidet P-40) for 30 min at room temperature using 20 μL reactions in 384-well plates. Fluorescence anisotropy was measured in a LJL Biosystems Criterion fluorimeter using a fluorescein filter set (excitation at 485nm and emission at 530 nm) and a G-factor of 0.8. Each point was measured three times and values were averaged from three independent experiments. K_ds were determined by fitting data to single-site binding curves using Kaleidagraph software.

Circular dichroism (CD) spectroscopy

CD spectra were measured using an Aviv model 62DS spectropolarimeter. Samples (50 μg/ml oligonucleotide) were prepared in 10 mM sodium phosphate pH 7.5, 100 mM NaCl and maintained at 4°C. Spectra were recorded from 320 to 210 nm using a 1 cm path length cuvette. The signal was averaged for five seconds at each wavelength and scans were repeated three times and averaged. For HIV TAR and analogs, 10 mM argininamide was added and for BIV TAR and analogs, a stoichiometric amount of peptide, 5μM, was added.

Native gel analyses and peptide binding gel shift assays

Oligonucleotides were heated to 85° C and slow cooled to 4°C (in 40 mM Hepes pH 7.5, 100 mM KCl, 1 mM $MgCl_2$, 0.5mM EDTA, 50 μg/ml yeast tRNA, 10% glycerol) and mobilities were characterized on native 20% polyacrylamide 1X TBE gels run at 600V for 20-24 h at 4°C. RNA-binding gel shifts with BTat 65-81 peptide were performed under similar conditions, with peptide at a saturating concentration (2 μM) incubated with oligonucleotides for 30 min at 4°C.

Results

Arginine binding to HIV TAR DNA/RNA hybrids

Early studies with the HIV Tat protein and ARM peptides showed that a DNA version of HIV TAR did not bind the protein specifically even though no 2' hydroxyl group was directly involved in binding, leading to the suggestion that ARM-RNA interactions may require an A-form helical geometry for recognition[13,14,33]. To better visualize how helix geometry might influence binding, we generated models of HIV TAR, based on an NMR complex with argininamide, in which the upper and lower stems were replaced by idealized A- or B-form helices (Fig. 1B). It is apparent that the distances between the O6 and N7 atoms of G26 and the phosphate of A22, which make the most critical contacts to arginine, are lengthened in the B-from model and cannot simultaneously hydrogen bond to the guanidinium group. It is less apparent whether differences between the major groove widths, between the orientations of the two helices, or between helical twists at individual base steps might also affect interactions either with the guanidinium group or the aliphatic portion of the arginine side chain, or how localized any important differences might be.

One way to effectively change helical conformation is to alter the 2' atom of the sugar ring, using a 2' hydroxyl group to favor the C3'-endo sugar pucker and resulting A-form geometry[6]. Indeed, some studies have suggested that introducing a single 2' hydroxyl into a DNA helix can substantially drive its conformation towards the A-form, although the extent to which the conformation propagates along a helix is less clear[34-38]. To identify regions in HIV TAR where the helical form is most critical for recognition,

we systematically introduced 2' hydroxyl groups into an all-DNA version of HIV TAR and measured their ability to rescue arginine binding, as monitored by salt-dependent elution from an L-arginine agarose column [23,39]. Because the 2' hydroxyl groups in an A-form helix project into the minor groove and cannot directly contact arginine or ARM peptides bound in the major groove, we reasoned that effects on binding should result largely from changes in local or global helical conformation. Strikingly, ribose substitution of the two base pairs immediately above and immediately below the bulge was sufficient to generate a hybrid (H2, Fig. 2) with an affinity for arginine ~2-fold higher than the all-RNA version of HIV TAR.

A variety of analogs in the upper stem (Fig. 2A), indicates the particular importance of ribose at the base pair immediately above the bulge. Removing ribose groups from both upper stem base pairs in the H2 context (H7) reduces binding affinity ~250-fold relative to H2, but is still ~15-fold tighter than the all-DNA analog, reflecting the importance of ribonucleotides in the lower stem. Ribose substitution only at the G26•C39 base pair (H3) reduces affinity by ~10-fold relative to H2 whereas substitution at A27•U38 (H4) reduces affinity by ~60-fold. Much of the affinity of H4 can be restored by substituting ribose at an additional adjacent base pair (H8), perhaps suggesting a cooperative propagation of A-form helix structure towards the binding site. Substitutions made only at positions in the 5' or 3' strands (H5 and H6) showed similar affinity to substitutions at G26•C39, indicating that at least one 2' hydroxyl group positioned at the junction of the two stems is important but that no particular position is critical. These results are consistent with effects of single deoxyribose substitutions in HIV TAR RNA on Tat peptide binding [40].

Analogs in the lower stem (Fig. 2B) indicate that riboses at the two base pairs below the bulge contribute approximately equally to binding. Removing ribose groups from both base pairs in the H2 context (H15) reduces binding affinity ~70-fold relative to H2, whereas either single base pair substitution (at A22•U40 or G21•C41; H11 or H12) restores binding by only ~2-3-fold. There is a marked strand asymmetry at these base pairs, with riboses being much more important on the 5' strand (compare H13 and H14). Previous work has shown that deoxyribose substitutions at either G21 or A22 have only a small effect on Tat peptide binding to the RNA form [40], but the effects in the DNA context shown here are more marked, likely reflecting the importance of local A-form geometry in the stacked stems surrounding the binding site. The local nature of the effect is especially apparent in the lower stem as substituting ribose at an additional adjacent base pair shows little evidence for helix propagation (compare H12 and H16).

The effects of some ribose substitutions in the nonhelical regions (Fig. 2C) were unexpected, in particular showing a marked preference for deoxyribose in the bulge (compare H2 and H19). There is NMR evidence that U23, which participates in base triple formation and is the only bulge residue critical for arginine binding, may exist in a C2'endo conformation when bound to argininamide, but the observed effect on binding is not specific to U23 (compare H18 and H19)[18,40,41]. In the context of an all-ribose bulge, it also is slightly favorable to have riboses in the loop (compare H17 and H20), resulting in an affinity similar to the all-RNA version of HIV TAR.

CD of HIV TAR DNA/RNA hybrids

The binding data demonstrate the importance of 2' hydroxyls at the base pairs immediately flanking the HIV TAR bulge. To assess whether these ribose substitutions result in localized A-form geometries or whether the helix propagates further along the stems, we estimated the relative A- and B-form helical content by CD, which gives characteristic spectra for the two helical forms [42]. The CD spectrum of the all-RNA version of HIV TAR is characteristic of the A-form geometry, with strong positive ellipticity near 265 nm and relatively weak negative ellipticity near 240 nm, whereas the all-DNA version shows the characteristics of B-form geometry, with strong positive ellipticity near 280nm and little signal near 240 nm (Fig. 3A). DNA/RNA hybrids containing the important ribose base pairs in the lower (H7) or upper (H15) stems, or the tight binding H2 hybrid, containing substitutions in both stems, produced intermediate spectra. Based on the ratio of signals at 266 nm (A-form) and 284 nm (B-form), we estimate 8% A-form content for H7, 14% for H15, and 19% for H2, consistent with localized induction of A-form geometry near the arginine-binding region and little propagation beyond the sites of ribose substitution.

We next asked whether arginine binding might lead to propagation of the A-form geometry beyond the binding region. CD spectra recorded in the presence of 10 mM argininamide (Fig. 3B) are not grossly different than in the absence of argininamide, with estimated A-form contents of 7% for H7, 13% for H15, and 14% for H2. However, the CD difference spectra of the tight binding H2 hybrid shows a significant decrease in signal near 280 nm (Fig. 3C), very similar to that observed with TAR RNA and indicative of a change in base stacking that is not seen with the poor binding H7 or H15

hybrids[16]. An opposite change in the CD signal is observed with the all-DNA analog, possibly suggesting some nonspecific interactions. The results further suggest that the effects of the 2' hydroxyl groups on helix structure remain largely localized to the sites of substitution.

BIV Tat peptide binding to BIV TAR DNA/RNA hybrids

To examine the helix geometry requirements for a more extensive ARM protein-RNA interface, we performed a similar set of experiments with DNA/RNA hybrids based on the BIV Tat-TAR complex. This complex also involves a specific arginine-RNA interaction that is virtually identical to that in the HIV complex, but it is presented in the context of a β-hairpin peptide along with seven other required amino acids that form an interface comparable in surface area to those seen in DNA-protein complexes [28,43]. To visualize the possible consequences of altering the helix geometry on BIV TAR recognition, we generated models, based on an NMR complex, in which the upper and lower stems surrounding the two single-nucleotide bulges (Fig. 1A) were replaced by idealized A- or B-form helices (Fig. 1C). By considering just two interactions - the Arg73:G11 contact analogous to that in HIV TAR and the Thr72:C23 phosphate contact - it is apparent that a major reorientation of the β-hairpin would be required to maintain important contacts at both ends of the upper stem in a B-form conformation. The wider major groove of a B-form helix would poorly accommodate the width of the β-hairpin, and numerous interactions to RNA base and backbone atoms, with their positions altered by differences in base pair tilt and displacement from the helix axis, would be disrupted. Given the diverse and flexible nature of ARM peptides and the preponderance of long,

flexible arginine side chains, it is of particular interest to know whether the BIV Tat domain can adapt to localized, or extensive, changes in helical geometry[44].

BIV Tat peptide binding affinities were measured for a series of BIV TAR DNA/RNA hybrids using affinity chromatography and fluorescence anisotropy assays, which are in good agreement. The all-DNA version of BIV TAR binds 1000-fold more weakly than the RNA version whereas an analog with three ribose base pairs in the lower stem and all-ribose in the upper stem where the peptide binds (B1) has a similar affinity as the RNA (Fig. 4A). In examining the upper stem requirements (Fig. 4A), we found that removing all 2' hydroxyls (B7) reduced binding to the level of DNA, while substituting riboses only at base pairs G11•C25 and A13•U24, generating an analog (B10) similar to the tight arginine-binding H2 HIV TAR hybrid, bound more tightly although still with 100-fold lower affinity than the RNA. However, adding just one more ribose base pair at the top of the stem (B9) brought affinity to within a factor of 8 of the RNA, and adding one additional ribose pair to the lower stem (B3) brought affinity to within a factor of 2 of the RNA. Surprisingly, adding riboses to base pairs G14•C23 and C15•G22, both of which are involved in specific peptide contacts, had less of an effect than substitution of the upper base pair (compare B2 and B3). Thus, the upper stem 2' hydroxyl requirements for BIV TAR are nearly identical to those of HIV TAR, with the addition of riboses to the top base pair and to one additional base pair in the lower stem, despite the very large differences in the interaction surfaces (Fig. 1). The importance of ribose at the top base pair in this hybrid context may be related to the requirement for having a Watson-Crick base pair at this position in the RNA context, and the extra base pair in the lower stem related to stem sequence requirements [28].

Despite the remarkably similar ribose requirements for arginine binding to HIV TAR and BIV Tat peptide binding to BIV TAR, the two show rather different local sensitivities to ribose substitution in the upper stem. Most notably, the G26•C39 base pair that hydrogen bonds to arginine in HIV TAR is ~6-fold more sensitive to substitution than the A27•U38 base pair that participates in the base triple (compare H3 and H4 in Fig. 2A) whereas the converse is true for the corresponding G11•C25 and A13•U24 base pairs in BIV TAR (compare B4 and B5 in Fig. 4A). Furthermore, peptide binding to BIV TAR shows a marked strand asymmetry, with a strong preference for ribose in the 3' strand (Fig 4A; B11 and B12) versus little preference in HIV TAR (Fig. 2A; H5 and H6).

In the BIV TAR lower stem, removing all 2' hydroxyls is less detrimental than in the upper stem, but still reduces binding by ~60-fold (Fig, 4B; B1 and B15). Ribose substitutions at two or three of the lower stem base pairs restore most of the binding affinity (Fig. 4B; B13, B14, B16). There is no apparent strand preference for 2' hydroxyls in the lower stem of BIV TAR (Fig. 4B; B17 and B18), unlike the preference for the 5' strand in HIV TAR (Fig. 2B; H13 and H14). Thus, as with the upper stem, the overall regions affected by ribose substitution are similar in the two cases but the local sensitivities are quite different.

In the bulge region of BIV TAR, the presence of 2' hydroxyls is slightly unfavorable for binding (Fig. 4C; B21 and B22), as also observed with HIV TAR (Fig. 2C). BIV TAR analogs with 2' deoxyuridine substitutions also show tighter binding (data not shown), suggesting that the effect results from the 2' hydroxyl group and not the absence of the thymine 5-methyl group. As with HIV TAR, NMR experiments also indicate that the bulge nucleotides exist primarily in the C2' endo conformation[30].

CD of BIV TAR DNA/RNA hybrids

It is interesting that the G14•C23 and C15•G22 base pairs in the upper stem, which are critical for BIV Tat peptide binding, show little effect of ribose substitution whereas the U16•A21 pair at the top of the stem shows a large effect, thereby allowing the minimally-substituted B3 hybrid (Fig. 4A) to bind with nearly the same affinity as BIV TAR RNA. We used CD to assess whether the upper stem might be completely switched to the A-form geometry in this hybrid, either with or without peptide bound, or whether some localized B-form helix in the middle of the binding site might be accommodated by the peptide.

Like HIV TAR, the spectra of the RNA and DNA versions of BIV TAR are characteristic of A- and B-form structure (Fig. 5A). The B20 hybrid, with three ribose pairs in the upper stem, has 35% A-form content, the B15 hybrid, with an all-ribose upper stem, has 47%, and the tight binding B3 hybrid, with additional substitutions in the lower stem, has 65%. Thus, it appears that the A-form helix does not propagate across the G14•C23 and C15•G22 base pairs, since substitution with ribonucleotides increases A-form content proportionally despite having flanking ribose base pairs in B20. The B7 hybrid, which contains three ribose pairs in the lower stem, has an unexpectedly high A-form content of 47%. We suspect that some propagation of A-form structure may occur in the lower stem or at the junction of the stems, which are better stacked in unbound BIV TAR than HIV TAR due to differences in the bulge configurations[29,30].

Upon BIV Tat peptide binding, a significant shift is observed in the spectra of the two upper stem hybrids (B20 and B15), with an increased signal near 265 nm (Figs. 5B

and 5C) suggestive of a switch to a more A-like conformation. The A-form content of hybrid B20 is estimated to increase from 35% to 55% and hybrid B15 from 47% to 60%, perhaps reflecting propagation of helix structure into the lower stem driven by the energy of binding. A relatively smaller change is seen with the tight binding B3 hybrid, with a calculated change in A-form content from 65% to 66%, suggesting that any required A-form structure, which does not seem to include the G14•C23 and C15•G22 base pairs, is preformed in this hybrid and thus little energy is lost upon binding. In contrast to the upper stem hybrids, BIV TAR RNA and DNA show little change in the CD spectra in the presence of peptide, whereas B7, with substitutions only in the lower stem, shows a slight decrease in signal near 265 nm (Figs. 5B and 5C).

Native gel analyses of BIV TAR DNA/RNA hybrids

Oligonucleotide mobility on native gels has been shown to be proportional to the fraction of B-form helix, in addition to chemical changes in the sugar, and thus can provide some indication of helical form[42,45,46]. The relative mobilities of the BIV TAR hybrids generally correlate well with the level of ribose substitution in both the upper (Fig. 6A) and lower (Fig. 6B) stem hybrids, supporting the premise that the effects of ribose substitution are relatively additive and localized, with little propagation of helical conformation. The similar mobilities of B4 and B6, which differ by one ribose substitution in the upper stem, or the slightly different mobilities of hybrids with similar levels of substitution (compare B4 and B5 in Fig. 6A, and B13 and B16 in Fig. 6B), however, suggests some propagation at the lower stem-bulge boundary, consistent with the CD results. The mobilities of hybrids with or without ribose substitution at the

G14•C23 and C15•G22 base pairs are substantially different (compare B1 and B3 in Fig. 6A, and B15 and B20 in Fig. 6B), consistent with the chemical substitutions and a localized difference in helix geometry.

Relative mobility shifts of BIV Tat peptide-hybrid complexes (Fig. 6C) are consistent with conformational changes in the lower stem inferred from CD as well as localized helix geometry. All hybrids containing three ribose pairs in the lower stem (B2, B1, B13, B16, B14, B15) form complexes of low mobility, consistent with an overall A-form structure despite different numbers of ribose substitutions in the upper stem. Thus, peptide binding may drive formation of A-form structure in the lower stem, even with no ribose present (hybrid B15). The tight binding B3 hybrid forms the fastest mobility complex, suggesting that the upper stem, and presumably the deoxyribose G14•C23 and C15•G22 pairs, is not entirely A-form. The B2 hybrid, which lacks ribose at the top base pair of the upper stem, also forms a complex with faster mobility than expected for a fully A-form upper stem, but unlike the B3 hybrid, this incomplete A-form helix is somewhat detrimental to binding.

Discussion

We have defined the minimal 2' hydroxyl group requirements that support specific recognition of HIV TAR by arginine and BIV TAR by a BIV Tat peptide. In both cases, we hypothesize that the hydroxyl groups create an appropriate A-form geometry for major groove binding and thus serve an indirect role in recognition. Surprisingly, the helical requirements appear highly localized to the binding regions, and the effects of ribose substitution on helix structure seem largely restricted to the altered base pairs. Furthermore, in the case of BIV TAR, it seems that two essential base pairs in the heart of the binding site need not be in an A-form conformation and we surmise that the BIV ARM peptide is able to adapt to a more B-like conformation. Given the localized nature of the effects, we examine the possible consequences of changes in helix geometry on individual contacts observed in the NMR structures of the two complexes.

HIV TAR-arginine interactions

Ribose sugars at four base pair positions in HIV TAR (Fig. 2A, H2) are both necessary and sufficient for arginine binding, and the requirement for each can be rationalized based on structural data. First, the NMR structures of TAR-argininamide complexes position the guanidinium group of arginine to hydrogen bond to the G26 base and simultaneously to make electrostatic contacts to the A22 and U23 phosphates (Fig. 1A)[18,20,21]. In the A-form geometry, the phosphates are substantially displaced along the z-axis relative to the center of a base pair (z_p) [47], effectively moving the phosphate of A22, and potentially of U23, towards the upper stem, and in closer proximity to G26 base

relative to its location in a B-form helix (Fig. 1B). Consistent with this requirement, removing 2' hydroxyl groups from the A22•U40 base pair in the H2 context (Fig. 2B, H12) reduces arginine binding affinity by ~15-fold. Second, removing hydroxyl groups from the G26•C39 base pair (Fig. 2A, H4), which reduces affinity by 30-fold, is likely to reposition G26 relative to the backbone phosphates and thereby disrupt the guanidinium binding interaction.

Third, it has been proposed that an overtwisting of the helix at the A22-G26 step helps creates a pocket in which arginine can stack between A22 and U23; the greater twist of a B-form geometry would be expected to alter the base overlap and stacking interactions. Removal of 2' hydroxyl groups from G21 and A22 in the lower stem decreases binding by 20-fold (Fig. 2B, H14), but has a minimal effect when removed from the bases on the opposite strand (Fig. 2B, H13). The G21 ribose may indirectly affect binding via the stacking of A22, since deoxyribose substitution of the G21•C41 base pair (Fig. 2B, H11) also reduces binding affinity by 10-fold. Thus, the G21/A22/arginine/U23 stacking network may be energetically coupled and may respond cooperatively to local changes in sugar geometry, reminiscent, on a local scale, of the cooperative changes in DNA proposed to take place in ethanol[34,35].

Fourth, a base triple between U23 and the A27•U38 base pair in the upper stem (Fig. 1A) is formed upon arginine binding[18,21]. Removing hydroxyl groups from the A27•U38 pair (Fig. 2A, H3) reduces binding affinity by 5-fold, perhaps by repositioning the base pair and disrupting the triple. The moderate effect on binding is consistent with other mutagenesis results of base triple residues[18,48]. Thus, the requirement for ribose at each of the four base pairs in HIV TAR has a reasonable structural explanation,

supporting the view that the effects of each ribose substitution are highly localized within or adjacent to the site.

BIV TAR-Tat peptide interactions

The BIV TAR-Tat complex provides an excellent model of ARM-RNA recognition, and this study reveals several new features of the interaction, described below, and also provides an interesting comparison to the HIV complex. The arginine-binding module of HIV TAR is conserved in the BIV complex and is very similar in structural detail, with Arg73 used to contact the G11 base and phosphate backbone in the BIV case (Fig. 1A) [28-30]. The ribose requirements at the arginine-binding site also appear to be similar, with the BIV interaction requiring 2' hydroxyls at the two base pairs immediately flanking the U10 bulge in both the upper and lower stems (Figs. 4A and 4B). The sequences of the lower stems differ, and studies of BIV TAR mutants showed that its sequence is important for binding, probably due to subtle stacking interactions [28]. Correspondingly, the BIV TAR-peptide interaction is enhanced by ribose substitution at one additional base pair in the lower stem (Fig. 4, B3 and B9). The only other difference in ribose requirements between the two complexes is the need for ribose at the top base pair of the upper stem in BIV TAR, as explained below.

The striking similarity in ribose requirements between HIV and BIV TAR seems remarkable given the large differences in binding surfaces for arginine and the BIV peptide (Fig. 1B and Fig. 1C). However, there are quantitative differences that likely reflect other constraints on the peptide complex. For example, the relative importance of the two upper stem base pairs is reversed: deoxy substitution of the A13•U24 base pair in

BIV TAR reduces affinity by 140-fold and substitution of G11•C25 reduces affinity by 10-fold (Fig. 4A, B4 and B5) whereas the converse if true for HIV TAR (Fig. 2A, H3 and H4). In addition, there are different strand preferences for the two complexes, with 2' hydroxyl groups being especially important on the 3' strand of the upper stem of BIV TAR (Fig. 4A, B11 and B12) and on the 5' strand of the lower stem of HIV TAR (Fig. 2B, H13 and H14). The preference for ribose in the 3' strand of the BIV site, as well as the requirement for ribose at the A13•U24 base pair, which is surprisingly insensitive to mutation, can be explained by the positioning of Gly74 between the base and ribose of U24 [27,29,30]. Gly74 is an essential residue that is deeply buried within the RNA-peptide interface and held rigidly to form the turn of the β-hairpin (Fig. 7A)[26]. A C2'-endo conformation at U24 would substantially reduce the size of the Gly74 binding pocket, forcing a reorientation of the β-hairpin within the groove to maintain other essential RNA contacts. Furthermore, the overtwisting of the A-form helix in the arginine-binding module increases the depth of the pocket, probably further enhancing binding specificity. Thus, the β-turn appears particularly well adapted to the A-form RNA helix geometry.

Our data suggest that some A-form helix may propagate into the lower stem upon peptide binding. It is observed that Arg77 interacts with the G9•C26 base pair at the top of the lower stem, forming a G9/guanidinium/U10 stacking arrangement, similar to the G21/A22/arginine/U23 arrangement in HIV TAR, that is further extended to include Arg70 stacking on U10 and Ile79 stacking against the C5/C6 surface of U10[30]. This extensive stacking network may help insulate BIV TAR from effects of deoxy substitution of the 5' strand of the lower stem, unlike the case for HIV TAR (see above), and could explain why peptide binding might lead to helix propagation.

Perhaps one of the most striking results is that 2' hydroxyl groups are not needed at two essential base pairs (G14•C23 and C15•G22) in the core of the BIV TAR binding site, located just above the arginine-binding module (Fig. 1A). CD and native gel data suggest that these pairs are not converted to A-form geometry despite flanking ribose pairs in the B3 hybrid (see Fig. 4A). In the BIV complex, Arg70 hydrogen bonds to the O6 and N7 groups of G14, Gly71 hydrogen bonds to G22, and Thr72 may hydrogen bond to C23 and stack on ribose. Because Arg70 does not make additional interactions to the phosphates or form other stacking interactions seen in the arginine-binding module, it should be largely insensitive to changes in helix geometry, as observed. The Gly71 and Thr72 interactions, which both contribute substantially to binding affinity, might be expected to be sensitive to helix geometry as both are largely buried on the inner strand of the β-hairpin and Thr72 contacts both a base and the backbone[26]. In these cases, the adaptability of the ARM and the flexible nature of the arginine side chains may allow substantial adjustments to accommodate any changes in helix geometry[10,11,44,49].

Finally, the sensitivity of BIV TAR to deoxy substitutions in the top base pair of the upper stem (Fig. 4A, B10) may reflect fraying of the upper stem at a putative B-form/A-form junction immediately adjacent to the loop. Mutagenesis of BIV TAR RNA has shown that having a Watson-Crick base pair at this position is important, but not its identity [28]. Introduction of such junctions generally lowers the thermal stability of an RNA, consistent the slightly larger effect observed in a fluorescence anisotropy assay at 25°C than in an affinity chromatography assay at 4°C (Fig. 4A, B10)[50,51]. Furthermore, the loss of affinity can largely be suppressed by ribose substitutions at G14•C23 and C15•G22 (Fig. 4A, B2 and B10) despite the lack of a requirement for 2' hydroxyls at

those pairs (see above). Effects on duplex stability clearly can indirectly influence a protein interaction.

Implications for other RNA-protein interactions

Both the requirements for specific A-form geometries and the mosaic character of a helical binding site undoubtedly apply to other RNA-protein complexes. The arginine-binding module discussed here is found in two aptamers that bind the HIV Rev peptide and in the HTLV RxRE RNA element and likely will show geometric requirements similar to those of HIV and BIV TAR[22,24,52,53]. The bacteriophage λ and P22 N peptide-box B RNA and RSG-1.2 peptide-RRE RNA complexes all utilize an ARM in which alanine is seen to interact in a pocket defined by the 2' and 3' carbons of the ribose sugar and the C5 and C6 carbons of a pyrimidine base (Fig. 7B)[54-56]. The confluence of hydrophobic groups is a consequence of the C3' endo ribose conformation and is the same pocket occupied by Gly74 of the BIV Tat peptide, suggesting that this may be a relatively common recognition feature in the RNA major groove. Another interaction likely to depend on helix geometry uses the arginine guanidinium group to make electrostatic contacts and/or hydrogen bonds to the oxygens of consecutive phosphates (Fig. 7C). These types of multivalent 'arginine fork' interactions[15], which can distinguish helix geometry because the distance between neighboring phosphate oxygens is ~1.4 Å shorter in the A-form than the B-form[6], have been observed in several ARM complexes[24,55,57]. On the other hand, it can be inferred that interactions made only to a base, such as an Arg-guanine interaction, may allow peptides or proteins to utilize particularly flexible side chains and electrostatic complementarity to adapt to different,

and possibly local, helix geometries[44,58]. The ARM may be especially well-suited to recognize a variety of helical contexts, including B-form helices found in structured single-stranded DNA elements [SGL and ADF, in preparation].

Studies of other RNA-protein interactions using 2' deoxy substitutions further highlight the importance of helix geometry. In the *E. coli* $tRNA^{ala}$•synthetase complex, where recognition occurs in the minor groove, and the MS2 coat protein-operator complex, where stem and loop sequences are recognized, removal of 2' hydroxyl groups believed to directly hydrogen bond to the protein reduced binding affinity by as much as 20-fold[59,60]. However, DNA analogs containing only the critical ribose groups still showed 20-fold reduced affinity, implying that helical conformation probably also is important. The double-stranded RNA-binding domain (dsRBD) provides the most extreme example of helix-specific recognition. Each modular dsRBD recognizes about one helical turn of RNA and makes a set of hydrogen bonding interactions to 2' hydroxyls and phosphate oxygens that precisely span the major, and perhaps the minor, groove of an A-form helix [61,62]. dsRBDs do not bind hybrids composed of one RNA and one DNA strand, and a small number of deoxyribose substitutions can significantly affect binding [63]. Thus, as exemplified by these cases and the TAR complexes, indirect readout of helical conformation can be energetically as important for high affinity binding as base-specific interactions or recognition of other tertiary features of RNA structure. It will be interesting to test the importance of localized helical changes in other complexes and to determine what types of RNA conformational features can be effectively insulated from neighboring regions.

Acknowledgements

We thank Alex Ramirez for help with anisotropy assays, and members of the Frankel laboratory for helpful suggestions and comments on the manuscript. This work was supported by a Howard Hughes Medical Institutes predoctoral fellowship (S.G.L.) and by NIH grants GM47478 and AI29135 (A.D.F.).

References

1. Cheng, A., Chen, W., Fuhrmann, C. & Frankel, A. Recognition of nucleic acid bases and base-pairs by hydrogen bonding to amino acid side-chains. *JMB* **327**, 781-796 (2003).

2. Garvie, C. & Wolberger, C. Recognition of specific DNA sequences. *Molecular Cell* **8**, 937-946 (2001).

3. Nekludova, L. & Pabo, C.O. Distinctive DNA conformation with enlarged major groove is found in Zn-finger-DNA and other protein-DNA complexes. *Proc Natl Acad Sci U S A* **91**, 6948-6952 (1994).

4. Suzuki, M. & Yagi, N. An in-the-groove view of DNA structures in complexes with proteins. *JMB* **255**, 677-687 (1996).

5. Shi, Y. & Berg, J.M. Specific DNA-RNA hybrid binding by zinc finger proteins. *Science* **268**, 282-284 (1995).

6. Saenger, W. *Principles of nucleic acid structure*, (Springer-Verlag, 1984).

7. Weeks, K.M. & Crothers, D.M. Major groove accessibility of RNA. *Science* **261**, 1574-1577 (1993).

8. Schroeder, R., Waldsich, C. & Wank, H. Modulation of RNA function by aminoglycoside antibiotics. *EMBO* **19**, 1-9 (2000).

9. Frankel, A.D. Fitting peptides into the RNA world. *Curr. Op. Struc. Biol.* **10**, 332-340 (2000).

10. Patel, D.J. Adaptive recognition in RNA complexes with peptides and protein modules. *Curr. Opin. Struct. Biol.* **9**, 75-87 (1999).

11. Weiss, M. & Narayana, N. RNA recognition by arginine-rich peptide motifs. *Biopolymers* **48**, 167-180 (1999).

12. Xie, B., Wainberg, M. & Frankel, A. Replication of human immunodeficiency viruses engineered with heterologous Tat-transactivation response element interactions. *Journal of Virology* **77**, 1984-1991 (2003).

13. Churcher, M.J. *et al.* High affinity binding of TAR RNA by the human immunodeficiency virus type-1 tat protein requires base-pairs in the RNA stem and amino acid residues flanking the basic region. *J Mol Biol* **230**, 90-110 (1993).

14. Dingwall, C. *et al.* Human immunodeficiency virus 1 tat protein binds trans-activation-responsive region (TAR) RNA in vitro. *Proc Natl Acad Sci U S A* **86**, 6925-6929 (1989).

15. Calnan, B.J., Tidor, B., Biancalana, S., Hudson, D. & Frankel, A.D. Arginine-mediated RNA recognition: the arginine fork. *Science* **252**, 1167-1171 (1991).

16. Tan, R. & Frankel, A. Circular dichroism studies suggest that TAR RNA changes conformation upon specific binding of arginine or guanidine. *Biochemistry* **31**, 10288-10294 (1992).

17. Tao, J. & Frankel, A.D. Specific binding of arginine to TAR RNA. *Proc. Natl. Acad. Sci. USA* **89**, 2723-2726 (1992).

18. Puglisi, J.D., Tan, R., Calnan, B.J., Frankel, A.D. & Williamson, J.R. Conformation of the TAR RNA-arginine complex by NMR spectroscopy. *Science* **257**, 76-80 (1992).

19. Brodsky, A.S. & Williamson, J.R. Solution structure of the HIV-2 TAR-argininamide complex. *J. Mol. Biol.* **267**, 624-639 (1997).

20. Aboul-ela, F., Karn, J. & 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**, 313-332 (1995).

21. Long, K.S. & Crothers, D.M. Characterization of the solution conformations of unbound and Tat peptide-bound forms of HIV-1 TAR RNA. *Biochemistry* **38**, 10059-10069 (1999).

22. Baskerville, S., Zapp, M. & Ellington, A. Anti-Rex aptamers as mimics of the Rex-binding element. *Journal of Virology* **73**, 4962-4971 (1999).

23. Tao, J. & Frankel, A.D. Arginine-binding RNAs resembling TAR identified by in vitro selection. *Biochemistry* **35**, 2229-2238 (1996).

24. Ye, X., Gorin, A., Ellington, A.D. & Patel, D.J. Deep penetration of an alpha-helix into a widened RNA major groove in the HIV-1 rev peptide-RNA aptamer complex. *Nat. Struct. Biol.* **3**, 1026-1033 (1996).

25. Ellington, A., Leclerc, F. & Cedergren, R. An RNA groove. *Nature Structural Biology* **3**, 981-984 (1996).

26. Chen, L. & Frankel, A.D. A peptide interaction in the major groove of RNA resembles protein interactions in the minor groove of DNA. *Proc. Natl. Acad. Sci. USA* **92**, 5077-5081 (1995).

27. Chen, L. & Frankel, A.D. An RNA-binding peptide from bovine immunodeficiency virus Tat protein recognizes an unusual RNA structure. *Biochemistry* **33**, 2708-2715 (1994).

28. Smith, C.A., Crotty, S., Harada, Y. & Frankel, A.D. Altering the context of an RNA bulge switches the binding specificities of two viral Tat proteins. *Biochemistry* **37**, 10808-10814 (1998).

29. Puglisi, J.D., Chen, L., Blanchard, S. & Frankel, A.D. Solution structure of a bovine immunodeficiency virus Tat-TAR peptide-RNA complex. *Science* **270**, 1200-1203 (1995).

30. Ye, X., Kumar, R.A. & Patel, D.J. Molecular recognition in the bovine immunodeficiency virus Tat peptide-TAR RNA complex. *Chem. Biol.* **2**, 827-840 (1995).

31. Macke, T. & Case, D. Modeling unusual nucleic acid structures. in *Molecular modeling of nucleic acids* (eds. Leontes, N. & SantaLucia, J.) (American Chemical Society, 1998).

32. Darby, N. & Creighton, T. Disulfide bonds in protein folding and stability. in *Methods in Molecular Biology*, Vol. 40 219-245 (Humana Press Inc., 1995).

33. Barnett, R.W., Delling, U., Kuperman, R., Sonenberg, N. & Sumner- Smith, M. Rotational symmetry in ribonucleotide strand requirements for binding of HIV-1 Tat protein to TAR RNA. *Nucleic Acids Res* **21**, 151-154 (1993).

34. Ban, C., Ramakrishnan, B. & Sundaralingam, M. A single 2'-hydroxyl group converts B-DNA to A-DNA. Crystal structure of the DNA-RNA chimeric decamer duplex d(CCGGC)r(G)d(CCGG) with a novel intermolecular G-C base-paired quadruplet. *J Mol Biol* **236**, 275-285 (1994).

35. Ivanov, V., Minchenkova, L., Minyat, E., Frank-Kamenetdkii, M. & Schyolkina, A. The B to A transition of DNA in solution. *JMB* **87**, 817-833 (1974).

36. Ivanov, V. & Krylov, D. A-DNA in solution as studied by diverse approaches. in *Methods in Enzymology*, Vol. 211 111-127 (Academic Press Inc., 1992).

37. Nishizaki, T. *et al.* Solution structures of DNA duplexes containing a DNA x RNA hybrid region, d(GG)r(AGAU)d(GAC) x d(GTCATCTCC) and d(GGAGA)r(UGAC) x d(GTCATCTCC). *Biochemistry* **35**, 4016-4025 (1996).

38. Salazar, M., Federoff, O., Miller, J., Ribiero, S. & Reid, B. The DNA strand in DNA•RNA hybrid duplexes is neither B-form nor A-form in solution. *Biochemistry* **32**, 4207-4215 (1993).

39. Tao, J., Chen, L. & Frankel, A.D. Dissection of the proposed base triple in human immunodeficiency virus TAR RNA indicates the importance of the Hoogsteen interaction. *Biochemistry* **36**, 3491-3495 (1997).

40. Hamy, F. *et al.* Hydrogen-bonding contacts in the major groove are required for human immunodeficiency virus type-1 tat protein recognition of TAR RNA. *J. Mol. Biol.* **230**, 111-123 (1993).

41. Sumner-Smith, M. *et al.* Critical chemical features in trans-acting-responsive RNA are required for interaction with human immunodeficiency virus type 1 Tat protein. *J Virol* **65**, 5196-5202 (1991).

42. Ratmeyer, L., Vinayak, R., Zhong, Y.Y., Zon, G. & Wilson, W.D. Sequence specific thermodynamic and structural properties for DNA.RNA duplexes. *Biochemistry* **33**, 5298-5304 (1994).

43. Nadassy, K., Wodak, S. & Janin, J. Structural features of protein-nucleic acid recognition sites. *Biochemistry* **38**, 1999-2017 (1999).

44. Wilkinson, T., Botuyan, M., Kaplan, B., Rossi, J. & Chen, Y. Arginine side-chain dynamics in the HIV-1 Rev-RRE complex. *JMB* **303**, 515-529 (2000).

45. Bhattacharyya, A., Murchie, A.I. & Lilley, D.M. RNA bulges and the helical periodicity of double-stranded RNA. *Nature* **343**, 484-487 (1990).

46. Roberts, R.W. & Crothers, D.M. Stability and properties of double and triple helices: dramatic effects of RNA or DNA backbone composition. *Science* **258**, 1463-1466 (1992).

47. Lu, X., Shakked, Z. & Olson, W. A-form conformational motifs in ligand-bound DNA structures. *JMB* **300**, 819-840 (2000).

48. Weeks, K.M. & Crothers, D.M. RNA recognition by Tat-derived peptides: interaction in the major groove? *Cell* **66**, 577-588 (1991).

49. Frankel, A.D. & Smith, C.A. Induced folding in RNA-protein recognition: more than a simple molecular handshake. *Cell* **92**, 149-151 (1998).

50. Soto, A., Gmeiner, W. & Marky, L. Energetic and conformational contributions to the stability of Okazaki fragments. *Biochemistry* **41**, 6842-6849 (2002).

51. Gyi, J., Conn, G., Lane, A. & Brown, T. Comparison of the thermodynamic stabilities and solution conformations of DNA•RNA hybrids containing purine-rich and pyrimidine-rich strands with DNA and RNA duplexes. *Biochemistry* **35**, 12538-12548 (1996).

52. Ye, X. *et al.* RNA architecture dictates the conformations of a bound peptide. *Chem. Biol.* **6**, 657-669 (1999).

53. Jiang, F. *et al.* Anchoring an extended HTLV-1 Rex peptide within an RNA major groove containing junctional base triples. *Structure* **7**, 1461-1472 (1999).

54. Cai, Z. *et al.* Solution structure of P22 transcriptional antitermination N peptide-box B RNA complex. *Nat. Struct. Biol.* **5**, 203-212 (1998).

55. Gosser, Y. *et al.* Peptide-triggered conformational switch in HIV-1 RRE RNA complexes. *Nat. Struct. Biol.* **8**, 146-150 (2001).

56. Legault, P., Li, J., Mogridge, J., Kay, L.E. & Greenblatt, J. NMR structure of the bacteriophage lambda N peptide/boxB RNA complex: recognition of a GNRA fold by an arginine-rich motif. *Cell* **93**, 289-299 (1998).

57. Battiste, J.L. *et al.* Alpha helix major groove recognition in an HIV-1 Rev peptide-RRE RNA complex. *Science* **273**, 1547-1551 (1996).

58. Garcia-Garcia, C. & Draper, D. Electrostatic interactions in a peptide-RNA complex. *Journal of Molecular Biology* **331**, 75-88 (2003).

59. Musier-Forsyth, K. & Schimmel, P. Functional contacts of a transfer RNA synthetase with 2'-hydroxyl groups in the RNA minor groove. *Nature* **357**, 513-515 (1992).

60. Baidya, N. & Uhlenbeck, O. The role of 2'-OH groups in an RNA-protein interaction. *Biochemistry* **34**, 12363-12368 (1995).

61. Ramos, A. *et al.* RNA recognition by a Staufen double-stranded RNA-binding domain. *EMBO* **19**, 997-1009 (2000).

62. Ryter, J.M. & Schultz, S.C. Molecular basis of double-stranded RNA-protein interactions: structure of a dsRNA-binding domain complexed with dsRNA. *EMBO J.* **17**, 7505-7513 (1998).

63. Bevilacqua, P.C. & Cech, T.R. Minor-groove recognition of double-stranded RNA by the double-stranded RNA-binding domain from the RNA-activated protein kinase PKR. *Biochemistry* **35**, 9983-9994 (1996).

Figure Legends

Figure 1. Models of HIV TAR-arginine and BIV TAR-Tat peptide complexes. (A) Schematic drawings of the HIV (left) and BIV (right) TAR complexes. Nucleotides in the RNAs and amino acids in the BIV Tat peptide important for binding are highlighted[26,27]. Phosphates whose ethylation strongly interferes with binding are indicated by black dots, and one with a moderate effect is in gray[15,27]. Hydrogen bonding and electrostatic interactions are indicated by arrows, and van der Waals interactions by dashed lines. (B) Overlapped models of B-form (left, yellow) and A-form (right, red) HIV TAR helices on the HIV TAR-argininamide complex, with argininamide shown in a ball-and-stick representation and the RNA backbone as ribbon (blue)[18]. Positions of the A22 phosphates are shown by spheres. (C) Overlapped models of B-form (left, yellow) and A-form (right, red) BIV TAR helices on the BIV TAR- Tat peptide complex, with the RNA and peptide backbone shown as ribbons (blue)[29]. The Arg73-G11 hydrogen bonding interaction is shown in green and the Thr72-C23 phosphate interaction in yellow. Positions of the C23 phosphates are shown by spheres.

Figure 2. Arginine-binding affinities of HIV TAR hybrids. Positions of ribose substitution are highlighted. K_{rel} is the dissociation constant of each hybrid relative to that of HIV TAR RNA, as determined by affinity chromatography. Substitutions in (A) the upper stem, (B) the lower stem, and (C) the bulge and loop are shown.

Figure 3. CD spectra of HIV TAR RNA (●), HIV TAR DNA (■), H2 (◆), H7 (□), and H15 (⊞) in the (A) absence or (B) presence of 10 mM argininamide. (C) Difference spectra calculated by subtracting the spectra in the absence of argininamide from the corresponding spectra in the presence of argininamide.

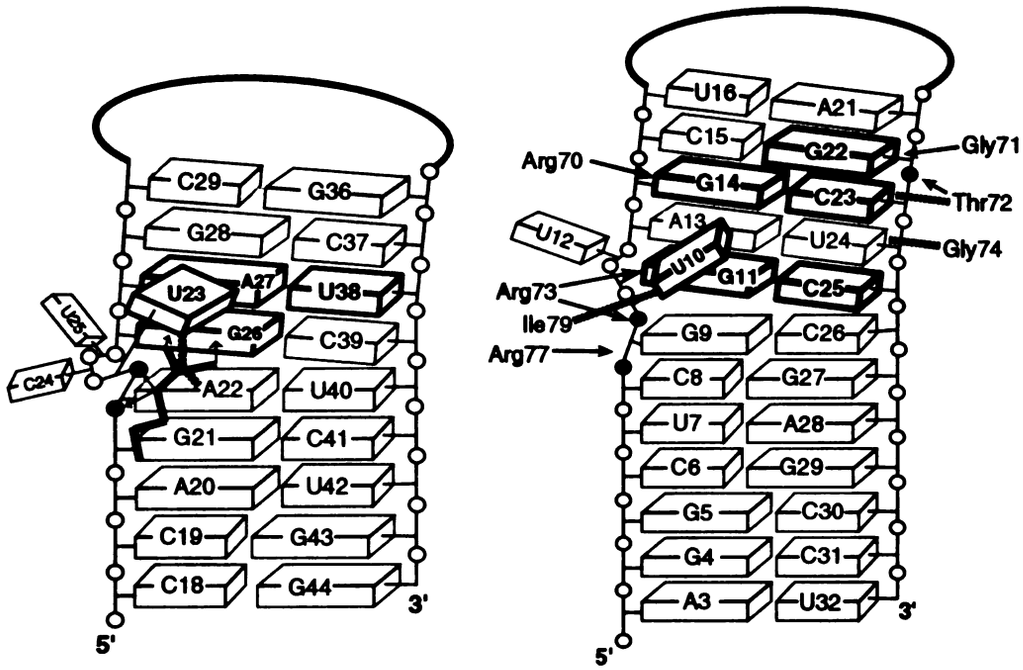
Figure 4. BIV Tat peptide-binding affinities of BIV TAR hybrids. Positions of ribose substitution are highlighted. K_{rel} is the dissociation constant of each hybrid relative to that of BIV TAR RNA, as determined by fluorescence anisotropy. For hybrids in which peptide binding could not be measured accurately by anisotropy (nd), or for some selected hybrids, K_{rel} was estimated by affinity chromatography (K_{rel} NaCl). Substitutions in (A) the upper stem, (B) the lower stem, and (C) the bulge are shown.

Figure 5. CD spectra of BIV TAR RNA (●), BIV TAR DNA (■), B15 (◆), B3 (○), B20 (□), and B7 (⊞) in the (A) absence and (B) presence of stoichiometric BIV Tat peptide. (C) Difference spectra calculated by subtracting the spectra in the absence of peptide from the corresponding spectra in the presence of peptide.

Figure 6. Native gel electrophoresis of BIV TAR hybrids in (A) the upper stem and (B) the lower stem. (C) Gel mobility shifts in the absence (-) and presence (+) of saturating (2 μM) BIV Tat peptide for the indicated hybrids.

Figure 7. Examples of ARM interactions sensitive to helical conformation. (A) In the BIV TAR-Tat peptide complex (accession) [Puglisi], the pocket formed by the U24 nucleotide (yellow) accommodates Gly74 and the turn of the β-hairpin peptide (red). (B) In the bacteriophage λ boxB RNA-N peptide complex (1FQF)[56], Ala3 of the peptide α-helix (red) occupies the hydrophobic pocket created by the C5 and C6 atoms of Cytosine5 and the C2' and C3' atoms of Cytosine4 (yellow). (C) In a Rev peptide-aptamer complex (1ULL) [24], bivalent hydrogen bonds are formed between nonbridging oxygen atoms at consecutive phosphates (yellow) and arginine residues (red). Arg38 interacts with the oxygens of C22 and U23 and Arg42 interacts with the oxygens of U23 and G24.

Figure 1A



SGPRPRGTRGKGRRIRR
65 70 71 72 73 74 76 77 79 81

Figure 1B

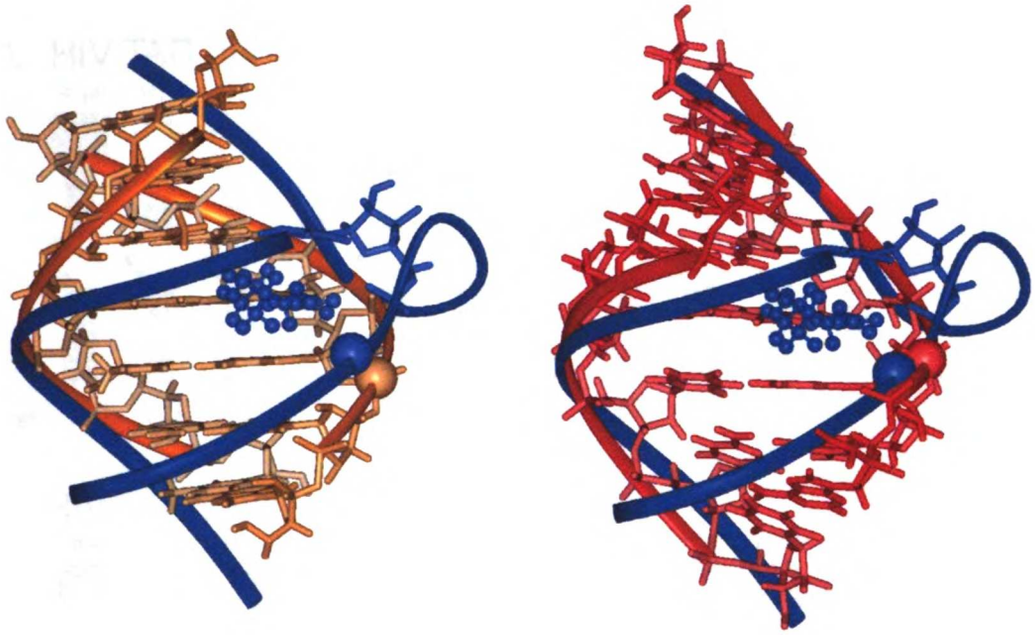
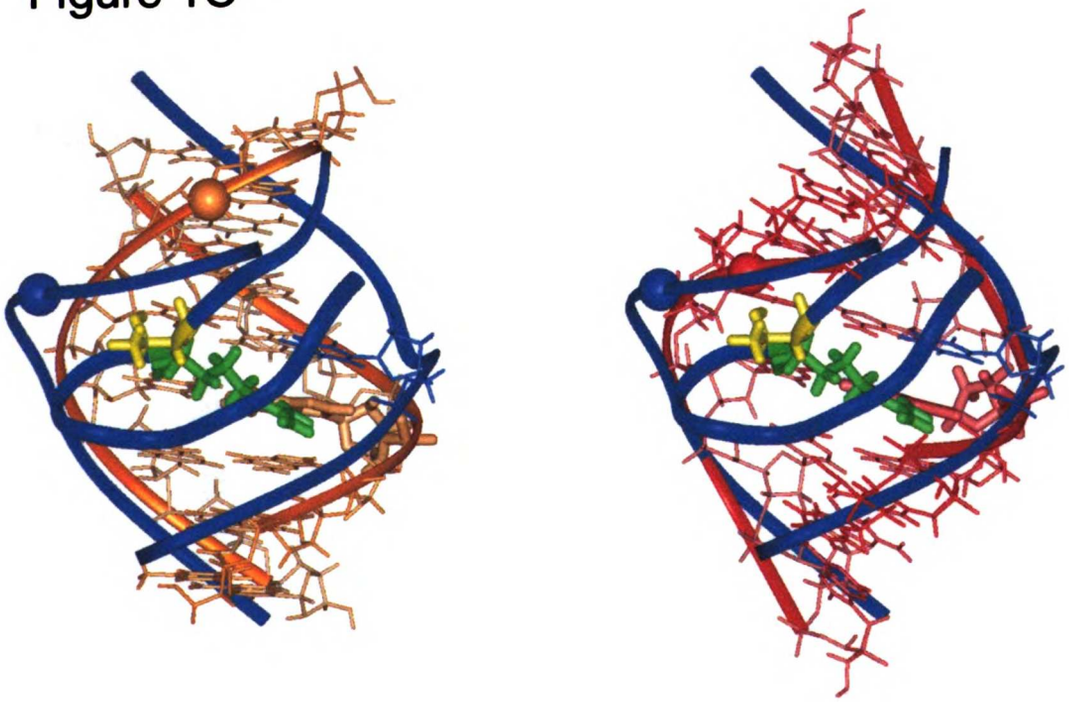
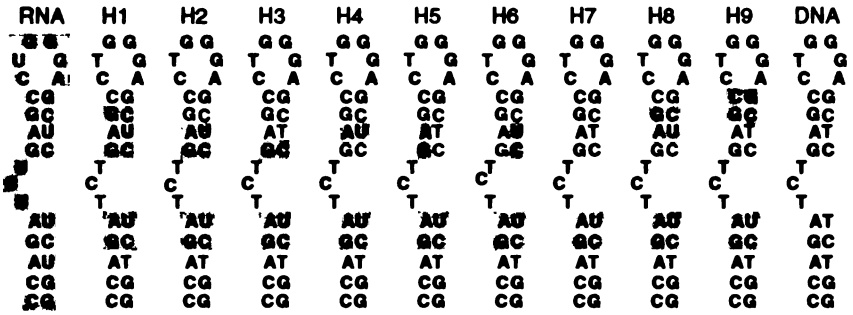


Figure 1C



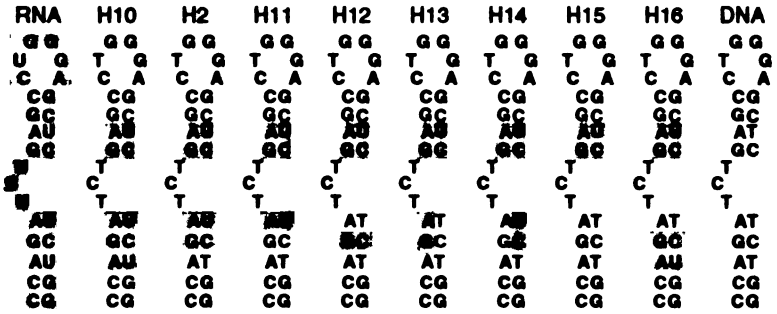
igure 2

A. HIV TAR upper stem



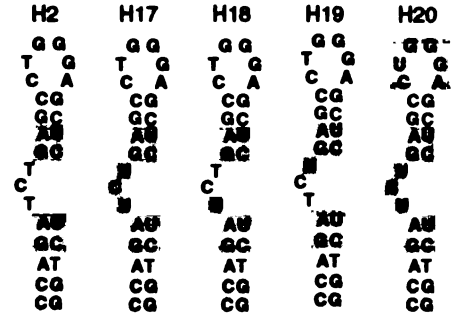
	RNA	H1	H2	H3	H4	H5	H6	H7	H8	H9	DNA
K_{rel}	1.0	1.0	0.5	5.0	31	6.2	5.0	120	5.0	76	1900

B. HIV TAR lower stem



	RNA	H10	H2	H11	H12	H13	H14	H15	H16	DNA
K_{rel}	1.0	0.7	0.5	10	15	2.3	20	34	6.7	1900

C. HIV TAR bulge/loop



	H2	H17	H18	H19	H20
K_{rel}	0.5	3.5	0.7	0.7	1.3

Figure 3

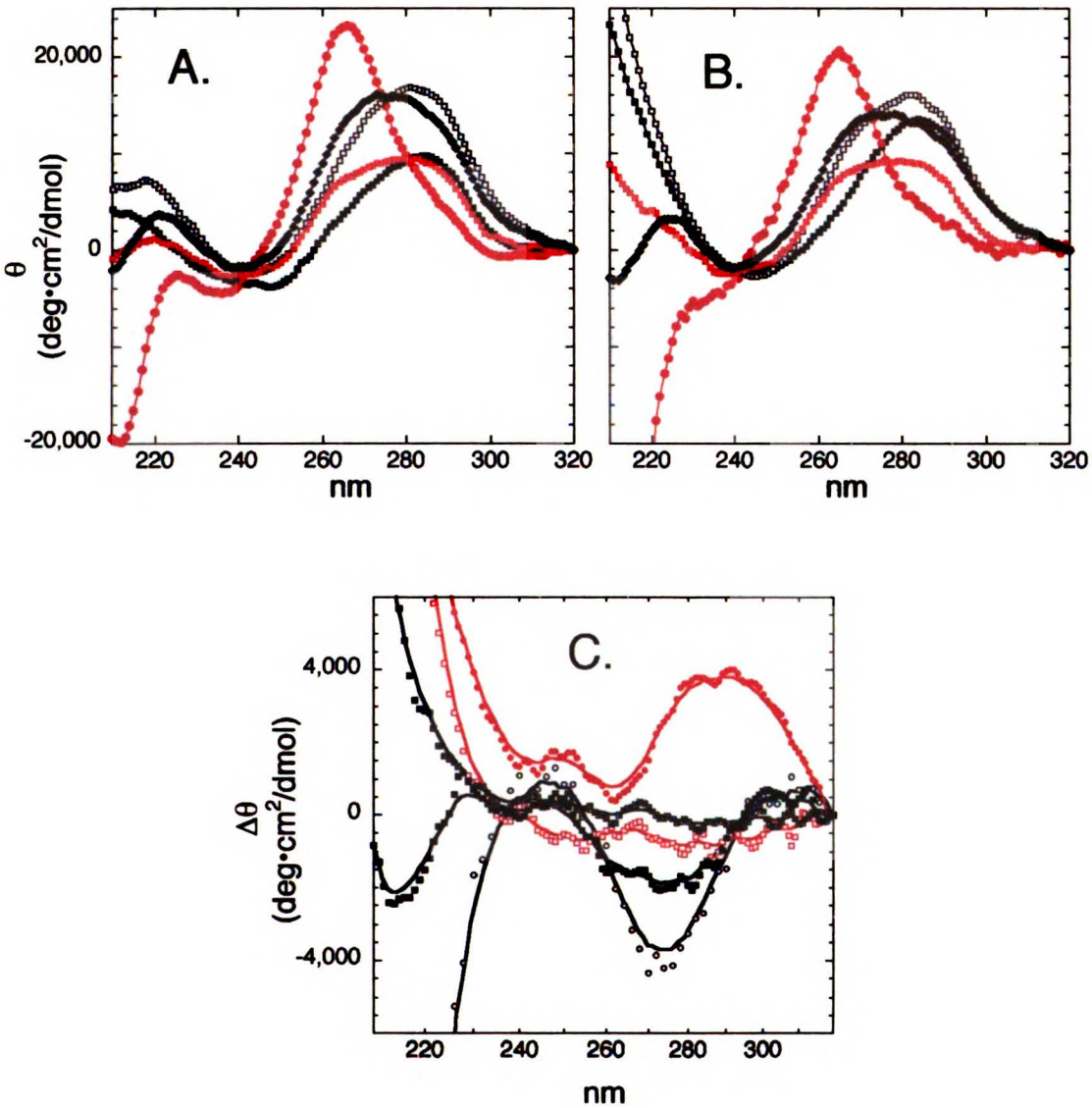


Figure 4

A. BIV TAR upper stem

	RNA	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10	B11	B12	DNA
K_{rel}	1.0	1.1	7.1	2.0	140	11	570	nd	nd	7.6	>200	31	nd	nd
K_{rel} NaCl	*1.0*				*420*		*1000*	*1400*	*140*		*110*	*6.9*	*500*	*1300*

B. BIV TAR lower stem

	RNA	B1	B13	B14	B15	B16	B17	B18	B19	B20	DNA
K_{rel}	1.0	1.1	2.5	7.3	60	3.0	1.5	2.5	2.0	170	nd
K_{rel} NaCl	*1.0*									*200*	*1300*

C. BIV TAR bulge

B21	B22
3.4	1.1

Figure 5

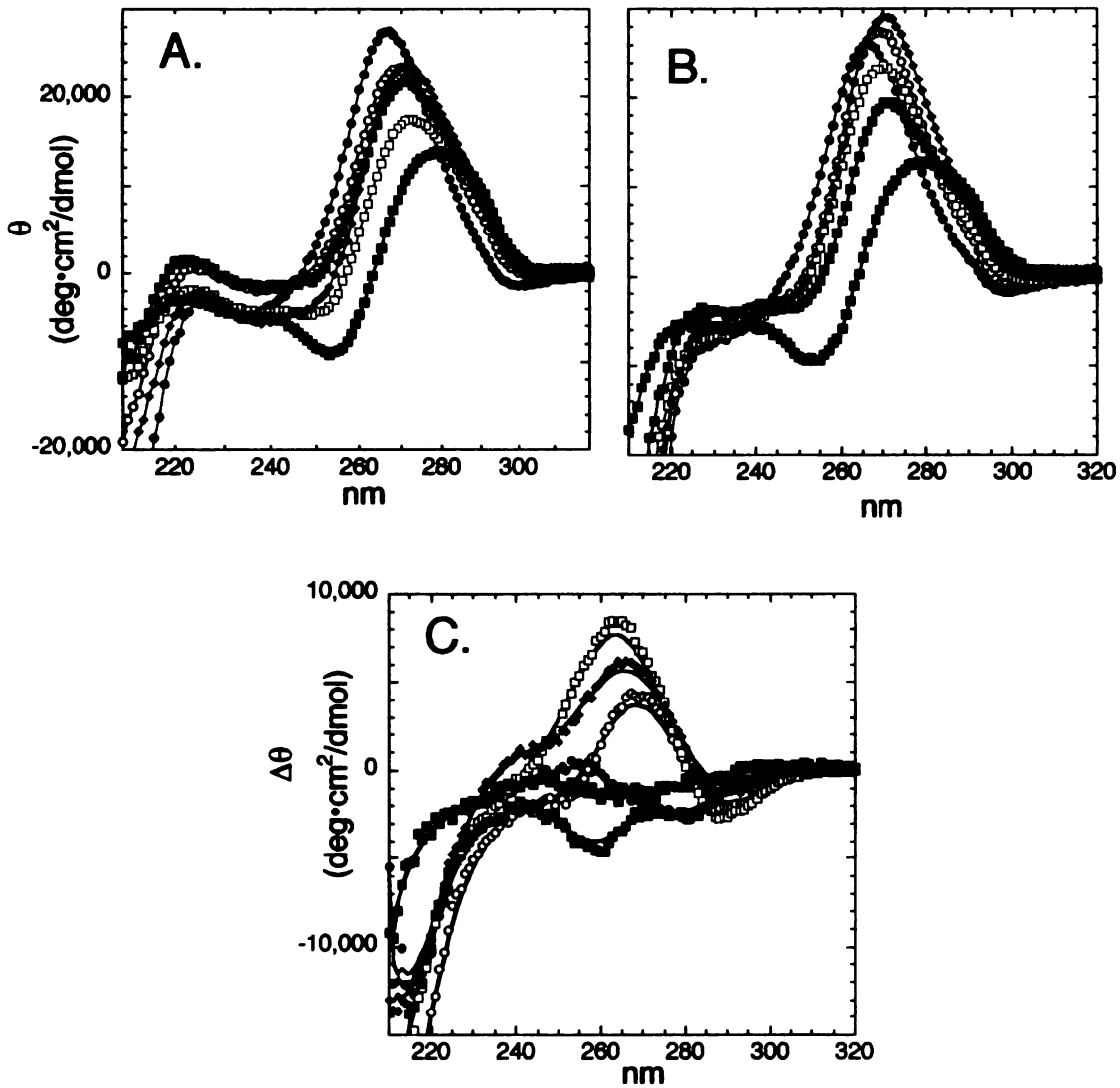

A.
B.
C.
20,000
0
-20,000
θ
(deg•cm²/dmol)
10,000
-10,000
Δθ
220
240
260
280
300
320
nm

Figure 6

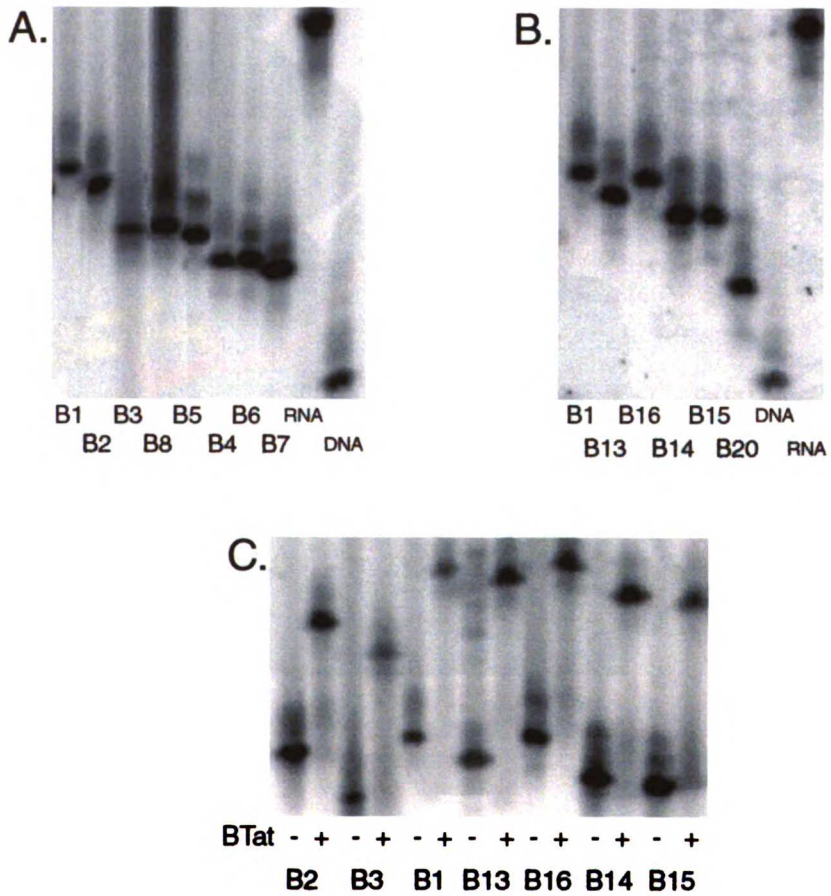
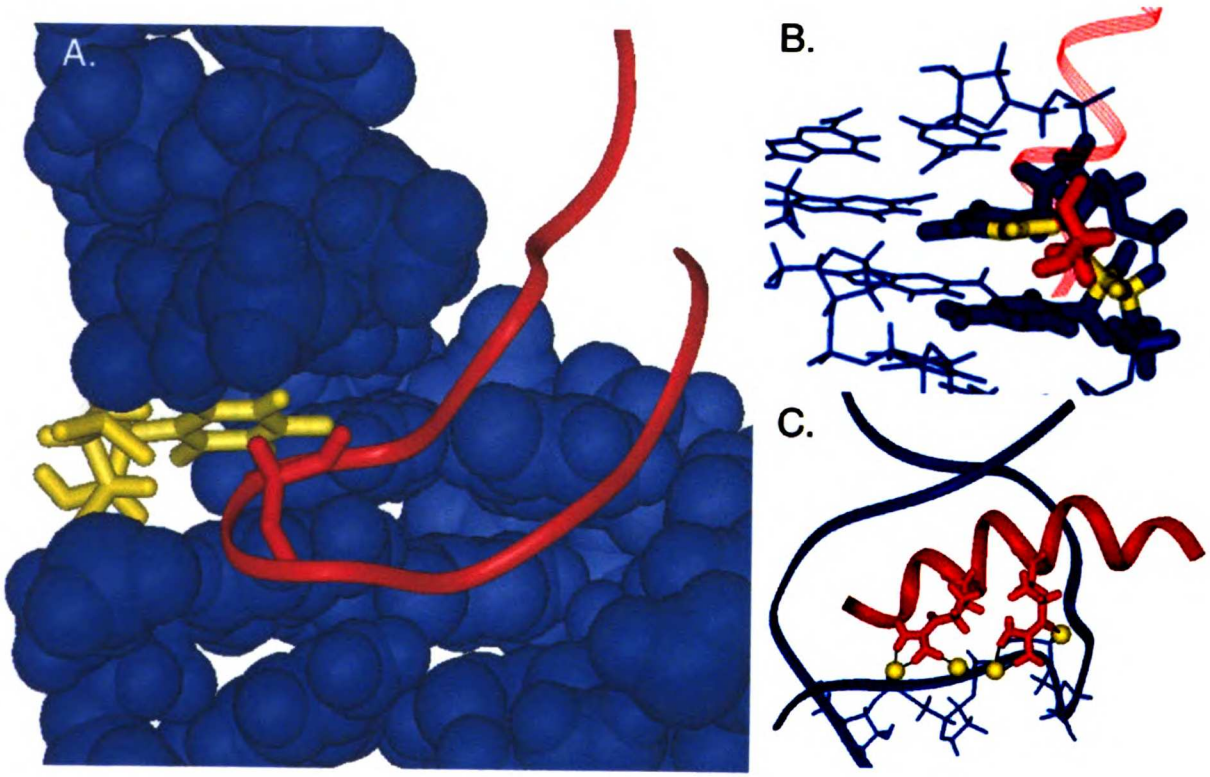

A.
B1 B3 B5 B6 RNA
B2 B8 B4 B7 DNA
B.
B1 B16 B15 DNA
B13 B14 B20 RNA
C.
BTat - + - + - + - + - + - + - +
B2 B3 B1 B13 B16 B14 B15

Figure 7



Chapter 3

A simple single-stranded DNA motif mediates high affinity protein recognition

A simple single-stranded DNA motif mediates high affinity protein recognition

Stephen G. Landt, Alejandro Ramirez, and Alan D. Frankel*

[1]Department of Biochemistry and Biophysics
University of California, San Francisco
San Francisco, CA 94143-2280

*Address correspondence to: Alan Frankel
Department of Biochemistry and Biophysics
UCSF
600 16th Street
San Francisco, CA 94143-2280

Telephone: 415-476-9994
FAX: 415-514-4112
e-mail: frankel@cgl.ucsf.edu

Abstract

Although there is much evidence that the specific recognition of single-stranded DNA (ssDNA) plays an important regulatory role in many phases of DNA function, little is known about how ssDNA tertiary structure can promote recognition. To address this we have used in vitro selection to identify ssDNA binders to the HIV Rev peptide, a protein motif adapted to structured RNA. The dominant feature to emerge from our selection was a two basepair stack, consisting of the non Watson-Crick G•T basepair on top of a CG basepair. We show that this motif is essential for high affinity binding in the context of multiple DNA folds, including a nearly double-stranded DNA hairpin and a branched 3-helix structure. We also show that it is specifically recognized by two arginine residues separated by one turn of a protein α-helix. This simple mode of binding explains the recognition of at least one critical biological regulatory ssDNA and we predict, because of its modularity and the affinity with which it is recognized, it will be found in many specific cellular ssDNA-protein interactions.

Introduction

Although DNA generally spends most of its lifetime in the double-stranded form, many of the most important events that happen to it occur while it is single-stranded. SsDNA (ssDNA) is a substrate for replication, recombination, repair, and transcription, and the potential for regulation of these processes via DNA conformation is great[1]. In many of these cases, it is to be expected that the recognition of specific ssDNA sequences by proteins will be essential for the proper fulfillment of ssDNA function. There are a handful of structurally characterized ssDNA-protein interactions involving the recognition of telomeric DNA sequences and sequences responsible for transcription factor binding[2-7]. In each of these examples, the DNA substrates assume little or no secondary or tertiary structure and instead are recognized by the many functional groups that are accessible in the extended conformation.

A general principle to emerge from the study of RNA-binding proteins is that RNA tertiary structure is essential for generating specificity. Particularly in light of recent progress in understanding protein•RNA interactions in the ribosome, it is clear that an enormous range of RNA structures are available for recognition[8]. These provide an array of unique backbone conformations, arrangements of base and backbone hydrogen bonding groups, surfaces for stacking interactions, and electrostatic distributions which can all contribute to creating specific protein binding sites. Despite this incredible conformational diversity, structural studies have been able to identify several common structural motifs which may be recognized in many contexts[9-13].

We expect that tertiary structures will also have an important place in ssDNA. Although these sorts of structures have proven difficult to identify because they are, by nature, transient, there are some biological examples of structured ssDNAs that are recognized by protein. These include an essential hairpin structure from the genome of the ssDNA phage N4 involved in initiating transcription and the intriguing role of a hairpin structure in the promoter of the enkephalin gene that is recognized by the CREB protein to activate transcription[14-17]. There is also the observation that an unexplainably high frequency of cruciform-like structures can be found in genomes[18,19]. These sorts of structures have the advantage of being relatively thermodynamically stable because of extensive basepairing, but also possess loops, and potentially internal bulges, non-Watson Crick basepairs and other features which might allow protein binding that is both tighter and more specific than for regular double-helical DNA.

It is our goal to begin to understand some of the basic structural principles by which structured ssDNA can be recognized. Our understanding of the recognition of linear ssDNA has been largely guided by the study of what are generally recognized as RNA-binding domains that also recognize ssDNA. Among the most common RNA-binding domains known are the RRM motif and the KH domain, and both have been shown to have the capacity recognize ssDNA with an affinity comparable to that with which they bind RNA[2,20]. Therefore, we felt that an RNA-binding motif optimized for structured RNA would provide a valuable model for the recognition of structured ssDNA.

The Arginine-Rich motif is a family of short, arginine-rich peptides which recognize folded RNA structures using multiple different protein folds and an obligatory induced fit mechanism[21]. Structurally characterized examples include the Tat peptides from HIV

and BIV, the HIV Rev peptide, the Rex peptide from HTLV and the N-peptides from the bacteriophages λ, P22, and HK022[22-31]. We have chosen the HIV Rev peptide as a target for the in vitro selection of ssDNA molecules that bind with high-affinity and specificity. A detailed biochemical characterization of two families of selected ssDNA molecules has revealed a variable balance between DNA structure and sequence in ssDNA recognition and most importantly, a small DNA motif that will likely be an element for ssDNA specificity in many biological contexts.

Results

In vitro selection of Rev-binding DNAs

The sequence of the 17 amino acid arginine-rich peptide from HIV-Rev, along with that of its high-affinity RNA-binding site (RRE-IIB) is shown in Figure 1a. One reason Rev was chosen is because it recognizes its natural RNA target in an α-helical conformation, by far the most common motif used for the recognition of double-stranded DNA[32]. As it has been shown that the α-helical form of the peptide is in equilibrium with other conformations, we carried out the selection with a peptide in which the α-helix is stabilized with a succinyl group at the N-terminus and a tail of 4 alanines at the C-terminus[33].

A DNA library of ~3×10^{14} sequences containing 50 randomized positions was screened for peptide binding. Since the peptide is extremely basic, we wanted to prevent the accumulation of DNA molecules that bind solely on the basis of electrostatics. To this end, we included a negative selection, eliminating sequences with high affinity for a peptide containing lysine substitutions at all the arginines in Rev. DNA molecules were cloned after 13 rounds of selection which included progressive increases in both the salt concentration in the binding reactions and the amount of competing lysine mutant peptide. Molecules from round 13 bound Rev with an affinity and specificity comparable to that seen for the RRE-IIB (see below) and were able to efficiently compete with the natural RNA for peptide binding (data not shown). Thus, a high-affinity, structure dependent RNA-binding peptide can also recognize ssDNA effectively.

Structure and sequence-specific peptide recognition

50 molecules were sequenced and a family, consisting of 27 members, all containing the pentamer TGTTC (Figure 1b), emerged. In 23 of these cases, this was accompanied by the presence of the tetramer AGCA, the complement of TGTTC with a G•T basepair at the second position of the predicted helix (Figure 1d). In most of these molecules, the AGC was provided by the last three positions of the 5' priming sequence. However, there were several cases where the only AGCA or the one most likely to be functional was composed entirely of randomized sequence. An additional seven sequences could be identified which contained single point changes in TGTTC and compensatory changes in the AGCA that restore Watson-Crick basepairing (Figure 1c). Of these 7, 5 occurred in the fourth position of the helix. Based on the predicted tertiary structures for these molecules, we believe that only the changes at the first position (TGTTC) represent functional covariations. Additionally, there were no instances in which the AGCA tetramer was preceded by a G which would pair with the C in TGTTC, suggesting that the unpaired state of this residue is important.

When these sequences were computationally folded, a 3-helix junction structure was a favored alternative in 22 of the 28 cases in which the AGCA/TGTT or a covariant pairing are present (Figure 1d). In this configuration, AGCA/TGTT comprise one helix and each of the 3 helices are tethered by linkers ranging from 0-4 bases, with no clear sequence conservation in these linkers. The validity of this fold is also supported by the observations that a nuclease boundary determination assay performed on several molecules delineates a minimal functional sequence that is consistent with the boundaries of the predicted 3-helix junction fold and that truncations introduced into 2 different

molecules which eliminate any one of the putative helices reduce binding at least 100-fold (in preparation).

We next wanted to determine the extent to which the sequence of the DNA determines specificity for the Rev peptide. With the 3-helix fold as a guide, we created a series of point mutants in 2 different TGTTC/AGCA molecules. Results for one sequence, A3, are detailed in Figure 2.

Binding to A3 and to an additional sequence, 29, occurs with an average affinity of ~1 nM, which is almost identical to the affinity Rev has for IIB RNA under these assay conditions. The selected DNA molecules bind at least 250-fold better than does DNA from the starting pool, compared to an ~100-fold difference between IIB and a non-specific RNA hairpin. Mutagenesis across the predicted structure shows that only the identities of positions within the conserved helix are essential for binding and that the G•T/CG basestep at positions 2 and 3 (TGTTC) is most important, with affinity losses of greater than 15-fold for any changes at these positions. Mutation of the non-Watson-Crick G•T basepair to either Watson-Crick basepair or a transversion to a T•G basepair all had large effects, demonstrating that the orientation as well as the identity of the basepair is important. More moderate affects are seen at the fourth position (TGTTC) while, consistent with the covariation, a mutant at the first basepair shows almost no loss of affinity. Interestingly, the identity of the apparently highly conserved unpaired C residue does not seem to be particularly important, since mutation to A in this molecule results in only a 6-fold loss of affinity and, in sequence 29, there is no loss (data not shown). This suggests that it is the unpaired character at this position which is recognized. Cumulatively, the identity of only three basepairs has even a moderate role in

determining specificity and, therefore, a large proportion of the binding specificity for this ssDNA family comes from recognition of the tertiary structure.

Because of the limited requirements for nucleotide identity, we wanted to more thoroughly understand where on the folded DNA the peptide was interacting. To do this, we carried out a series of chemical modification interference experiments using DMS to methylate the N7 position of guanines, hydrazine to eliminate pyrimidine bases, and ENU to ethylate non-bridging oxygens in the phosphate backbone. The results, shown in Figure 2c and summarized in 2d, are consistent with the mutagenesis. Methylation, a relatively subtle modification, only strongly affects recognition of the two G residues in the essential G•T//CG basestep, while the more disruptive hydrazine modification strongly affects all pyrimidines in the AGCA/TGTT helix. Furthermore, the affects of N7 modification by DMS on binding indicate that the helix is recognized in the major groove.

ENU modification affects recognition of phosphates on both sides of the conserved helix, but interferes much more at the final 2 residues on the 3' side of the helix and the unpaired C residue at the helical junction. Since this residue is only moderately affected by hydrazine or site-directed mutagenesis, it may be the backbone conformation at this position that is important, with the base identity serving to stabilize this conformation (or prevent an alternate one).

Finally, both DMS and hydrazine implicated residues in the helix 3' to TGTTC, although in both cases the interferences were demonstrably weaker. In experiments done on three other TGTTC family members, one additional helix generally shows a weak interference pattern, while the other is unaffected by

modification, although which one, relative to the conserved helix, is affected varies. This further supports the proposed 3-helix fold and is consistent with a structure in which one of the helices is oriented to allow non-sequence specific contacts, while the other is out of contact range with the peptide.

Sequence and structure-specific peptide recognition

Though the TGTTC/3-helix mode of recognition was the obvious winner in the selection, we also became interested in a second family, consisting of only 2 members. This family is characterized by two repeats of the G•T/CG basestep that comprise the most essential basepairs in the major class. In this case, analysis of deletions (data not shown) in one member of the smaller family (R28), defines a minimal region that is predicted to fold into a nine basepair stem, with a single-base bulge near the bottom and a four nucleotide loop (Fig 3a). Between the two family members, only the six basepairs immediately above the bulge are identical (Fig. 3b), suggesting that it is essentially double-stranded DNA that is recognized and that sequence identity should play a large role in determining specificity.

We determined the affinities of a series of mutants to identify essential DNA residues and the results are shown in Figure 3a. Mutations at five of the six conserved basepairs reduce binding by at least ~50-fold and mutation at the sixth reduces binding another 5-fold. No other basepairs or the loop sequence are important. The identity of the bulged residue also does not matter, although a 16-fold loss of binding occurred when it was deleted, indicating some role for a disruption of the double-helix.

As before, the G•T basepairs were examined with several mutants. The most striking feature for both pairs was the extreme loss of affinity when either G•T was mutated to G-C (affinities in both cases were too low to accurately measure) suggesting a dominant role for the T in defining the function of each G•T. Mutation of the G in either context also showed a large effect, but the transversion mutations behaved quite differently, with one showing a 140-fold loss of binding, while the other only reduced binding by 6-fold. The latter may be a case where the non-Watson Crick arrangement in the G•T basepair functions partly to affect the helix geometry, a role the transversion might also support, or Rev might simply be better able to adapt to the altered configuration in this transversion.

Modification interference verified the results of the mutagenesis. Every G residue and pyrimidine in the essential regions were extremely sensitive to modification, but no DMS effects were seen outside this region and hydrazine effects were limited to the immediately adjacent basepairs. ENU modification also revealed a more extensive pattern of interference than seen for the TGTTC family, with approximately nine positions showing at least some sensitivity to ethylation. This is consistent with measurements of the dependence of binding affinity on cation concentration, which showed that ~6 ions are released upon Rev binding to R28, compared to the release of ~5 ions upon binding to RRE IIB RNA and only 4 upon binding to TGTTC DNA (data not shown). Additionally, the oxygens at the base of the 5' strand that are most sensitive to ethylation are separated from the only sensitive oxygens on the 3' strand by ~1/2 helical turn such that they are directly across the major groove from one another, indicating that the peptide is able to make a substantial number of backbone contacts to both sides of the major groove in the region near the bulge.

Because Rev in particular and arginine-rich motif peptides in general appear so well adapted to the deep major groove of the A-form RNA helix, it was somewhat surprising to find that a DNA sequence bound by Rev with high affinity would consist of an almost uninterrupted double helix. This raised the possibility that R28 might adopt a more A-like conformation in the presence of Rev. To address this, CD spectra were taken of the R28 hairpin as well as an RNA version of R28 and of DNA and RNA versions with Watson-Crick basepairs substituted for each G•T. The results (Figure 4) show that R28 assumes a conformation in the absence of Rev that is generally B-like and clearly different from the RNA version of the molecule. When Rev is added, R28 undergoes a conformational change, revealed by the peak in the difference spectrum centered around 280nm, to a DNA structure that is indistinguishable by CD from the B-form Watson-Crick helix. It therefore appears that R28 is recognized in the major groove of its B-form DNA helix by a peptide α-helix, and that this interaction may be more comparable to DNA recognition by canonical double-stranded DNA binding proteins than to recognition by most other single-stranded RNA or DNA binding proteins.

Diverse mechanisms of ssDNA recognition

To begin to understand how the same peptide can use two different mechanisms of recognition, structure-specific and sequence-specific, we used alanine scanning mutagenesis to identify peptide residues essential for binding to each DNA family. The results for both families are shown in Figure 5a. As expected, arginine is essential for recognition in each case, with mutation of a pair of arginines (38,42) reducing binding to the 3-helix DNA by > 50-fold, while mutation of four arginines (38.41.42.46) reduces

binding to R28 from 10- to 170-fold. The notable difference is the essential role for tryptophan 45 in the recognition of the 3-helix DNA, suggesting that this is a primary determinant of DNA structure recognition (see below, discussion). However, in the context of the helical R28 molecule, it appears that arginine residues alone are sufficient to encode all the determinants of high affinity recognition.

An additional mutation was also introduced at a non-essential position, arginine 43 to proline, and binding to each DNA family was reduced at least 30-fold. Since the proline sidechain disrupts α-helices, this verifies that Rev recognizes both DNA families as an α-helix. When essential peptide residues are arrayed on a helical wheel diagram (Figure 5b), a distribution consistent with α-helical binding is seen and it appears that essentially the same face of the helix has been selected for in both cases. This contrasts with Rev binding to IIB RNA, where essential residues are distributed all around the helix surface. The difference is probably rooted in the different geometries of the nucleic acid helices that are recognized. The A-form RNA helix has a deep major groove which can accommodate a much larger fraction of the surface area of an α-helix than the B-form DNA structure assumed by R28 and by some or all of the 3-helix junction binding site. Remarkably, despite this apparently reduced contact area, Rev can still bind to DNA with an affinity comparable to that for RNA.

Finally, we wanted to identify specific amino acid sidechains responsible for the recognition of important DNA features, particularly the essential G•T/CG basesteps. To achieve this, we measured the affinities of several of the peptide mutants for a series of mutants in the G•T/CG steps of R28, hoping to find amino acid positions that determine specificity only for certain parts of the DNA. Our results (Figure 6) show that mutation

of arginine 42 substantially reduces discrimination against the G37A and C38T/G45A mutations in the upper stem, while having no affect on recognition of the G•T/CG basestep in the lower stem. A similar effect is seen for arginine46, although the loss of specificity is greater for the C38/G45 mutation. A reciprocal outcome was obtained for the lower stem, where the arginine 38 mutation cannot discriminate against the G34A/C49T or G48A mutations, and mutating arginine 41 eliminates discrimination against G34A/C49T. However, both of these mutants can still discriminate against the upper stem mutation G37A. From this, it appears that the peptide uses a pair of arginines to recognize each G•T/CG basestep and that the Rev helix is oriented with the amino terminus toward the base of the stem and the carboxyl terminus toward the loop. Furthermore, in each case, the paired arginines are separated by approximately one turn of the α-helix, suggesting that this spacing is ideal for recognizing a G•T/CG DNA basestep.

Discussion

G•T/CG as a general feature of ssDNA recognition

From our in vitro selection, the G•T/CG basestep emerged as the conserved recognition feature, present three times in two independent families of DNA molecules. The sequence surrounding the G•T/CG varies and the overall DNA structural context in which it is recognized differs between the families, indicating that this motif is modular and should be able to function in many nucleic acid structural contexts. In fact, it also appears in a ssDNA molecule selected to bind another ARM peptide (data not shown), the BIV Tat peptide, which binds RNA in a β-hairpin conformation. Most interestingly, it appears twice in a naturally occurring ssDNA hairpin located in the promoter for the enkephalin gene and bound by the CREB protein [15].

In the enkephalin case, two sites that are imperfect matches to the consensus binding site for CREB are found in the promoter and are necessary for signal-dependent activation via CREB. It was hypothesized that the sites, which are nearly inverted repeats of one another, undergo a transition from a double-stranded to a cruciform conformation in which one of the hairpins is bound by CREB. This model was subsequently validated by in vitro binding assays and in vivo footprinting assays [16,17]. The essential difference between the hairpin form and the double-stranded form is the presence of the two G•T/CG repeats in the hairpin, of which one is absolutely necessary for CREB binding and transcriptional activation. CREB is member of the basic leucine zipper (bZIP) family of double-stranded DNA binding proteins, suggesting that the G•T/CG motif can function in the context of an independent nucleic acid binding motif.

The results of the peptide mutant/DNA mutant cross-interaction experiments indicate that this motif was selected because it forms a specific binding site for two arginine sidechains. In this respect and in its modularity, it is reminiscent of the small RNA module first characterized in HIV TAR. TAR consists of a GC/AU basestep above a bulged U residue[27]. Arginine recognizes the G residue through a pair of hydrogen bonds to its major groove face and also makes an electrostatic contact with the phosphate oxygen of the bulged U, which is shifted upward to participate in a base-triple with the AU basepair[27,34,35]. Subsequently, this module has been shown to be an essential component of several natural and in vitro selected binding sites for a variety of arginine-rich peptides [36,37].

Although we cannot accurately say how the G•T/CG steps are recognized, the strong electronegative character of the major groove face of both the G•T basepair and the G/T basestep, with a total of five potential hydrogen bond acceptors and no donors, allows several potential hydrogen bonding arrangements that are complementary to the electropositive arginine sidechain (Figure 7a)[38]. One arrangement involves hydrogen bonds between arginine and the O6 and N7 groups on the major groove face of the guanine and the O4 of the thymine in a G•T basepair (Figure 7b). This interaction would be especially discriminating because it specifies two bases with one sidechain and the three hydrogen bonds should also make it very stable. In fact, from a computational study that identified all possible hydrogen bonding arrangements between nucleic acid basepairs and amino acid sidechains, this was the only interaction found that was capable of forming three hydrogen bonds to span a basepair [39].

Arginine should also be able to recognize both bases in the G/T basestep by donating hydrogen bonds to the O6 or N7 positions of the guanine and the O4 of the thymine. An interaction that spans this basestep is consistent with the data in Figure 6 for the arginine 41 mutant. This mutant has a reduced ability to discriminate against both a mutation in the essential GC basepair in the lower stem and against a transversion of the adjacent T•G basepair to G•T (data not shown), but still discriminates when the G residue in the T•G basepair is mutated to A and the T is left intact. Finally, the G residue of the GC basepair may be recognized by a standard hydrogen bonding interaction between arginine and the O6 and N7 groups, such as is seen in the TAR:arginine complex and very commonly in protein:DNA complexes[39].

As in the TAR:arginine interaction, these hydrogen bonding interactions might also be coupled to electrostatic contacts or hydrogen bonds to the phosphate backbone. Our ethylation interference data shows that modification of the phosphate oxygens at both residues of the GT basestep strongly interferes with binding in all three instances, while the effects on the opposite strand are much smaller. The specificity of any of these interactions might be further augmented by the unique geometry of the G•T basepair, where the T residue is pushed forward into the major groove.

It should be pointed out that any of these interactions might theoretically occur alongside any other and, since two arginines are responsible for the recognition of each G•T/CG basestep, it is probable that most of the five hydrogen bond acceptors will be engaged in hydrogen bonds. Furthermore, the spacing between the arginine pairs is similar in all three cases (arginines 38 and 41 and arginines 42 and 46 for R28 and presumably arginines 38 and 42 for the TGTTC/3-helix family), a spacing that coincides

with approximately one turn of the α-helix. It has long been known that the α-helix is the most often used protein structure for recognizing the DNA major groove and that arginine is the most important amino acid for specifically contacting DNA, but the motif we have selected appears to form an ideal complement for these structures that is available only to ssDNA[32,39,40]. dsDNA does not form any basepairs like the G•T that contain strictly hydrogen bond acceptors and, while G•T basepairs are quite common in RNA, the depth and width of the major groove of the A-form helix does not allow ready access to this surface.

Whichever way the G•T/CG is recognized by the arginines in Rev, the same sequence in the enkephalin hairpin is probably recognized similarly by CREB. Not only is the basestep identical, but mutation of the T residue in the G•T basepair is also substantially more detrimental than mutation of the G, an effect that is also clearly seen for both repeats in R28 (Figure 3b)[15]. CREB also recognizes its double stranded binding site, and presumably the hairpin, as an α-helix. In this helix, there are three arginine residues-294,298, and 301-, each separated by roughly one helical turn, that are oriented towards the major groove and would be in reasonable proximity to the G•T/CG[41].

For either of these arginine pairs to contact the G•T/CG step, a reorientation of either the CREB helix or the DNA helix, relative to their positions in the double-stranded DNA complex, is needed. However, the hairpin configuration alters several conserved features of the consensus DNA site, so a rearrangement must take place for hairpin binding to occur. Furthermore, the G•T basepair has been associated with increased dynamics and flexibility in the DNA helix and there is evidence for bZIP helices showing flexibility to recognize variant binding sites[42-44]. Although little progress has been reported in

designing altered specificities for bZIP helices, it may be that this motif is one of the few cases where enough high affinity contacts can be made to compensate for the loss of multiple highly conserved contacts from the dsDNA complex.

Variable roles for sequence and structure in recognition

Although the TGTTC family contains only one copy of the G•T/CG motif, it is able to bind Rev as tightly as the R28 family, with two G•T/CG repeats. Since there is only a limited sequence dependence in the regions outside the G•T/CG basestep, it must be recognition of the DNA tertiary structure that compensates for the absence of the second G•T/CG. Computational folding along with deletion analysis suggests that a 3-helix junction structure is the relevant fold for this family.

This type of structure is common in large RNA molecules and, by virtue of having branched helices, is also reminiscent of the 4-helix junctions that are intermediates in DNA recombination and repair. For each of these classes, there are structurally well-characterized examples of proteins that recognize them and some recognition principles are known. The best studied case of recognition of an RNA 3-helix junction is the interaction between the ribosomal protein S15 and 16S rRNA[45,46]. In this example, it is a unique backbone conformation at the helical junction, caused by an unusual base triple, that is recognized by the protein. In general, the DNA junctions are recognized in a structure-specific manner by protein dimers or tetramers with electrostatic and hydrogen bonding complementarity to the backbones of the four helices[47-50]. In the case of branch migration proteins such as RuvA, these interactions are limited to the helical arms, while junction-resolving endonucleases also make contacts near the helix junction.

If we assume that arginines 38 and 42 of Rev are involved in recognizing the G•T/CG basestep in the TGTTC fold, then the primary peptide determinant of specificity for the 3-helix junction must be tryptophan 45. This residue is spaced about one helical turn from arginine 42 at the G•T/CG, and, if the peptide is oriented with its C-terminus towards the helical junction as the modification interference data suggests, the tryptophan should be at or very near the junction. This indicates that it is the junction itself that is probably recognized and it strongly suggests the possibility that the tryptophan may be stacking with one or two of the three helix ends or with the conserved unpaired C residue in this region.

Tryptophan stacking is known to be important in the binding of many non-specific ssDNA binding proteins to unstructured ssDNA and in the binding of the HIV nucleocapsid protein to single-stranded RNA and DNA[51,52]. Furthermore, tryptophan stacking can also help determine sequence specificity, as seen in the recognition of the boxB RNA hairpin by the arginine-rich N-protein of bacteriophage λ, where tryptophan caps an extended base stack by stacking on an adenine residue at the apex of the loop [23]. The base-amino acid overlap is extensive and establishes the specificity of the peptide for a purine at this position in the loop[53]. Tryptophan has also been shown in several cases to contribute to specific DNA recognition by making important hydrogen bonds to the DNA backbone[54,55]. In our case, the confluence of helix ends that is characteristic of branched helix structures provides an opportunity for high affinity stacking with at least one helix end and perhaps two that can be easily coupled to recognition of specific sequences in the accessible major groove of DNA.

Though tertiary structure is necessary for recognition of the TGTTC family, the R28 family shows that an essentially double-helical DNA molecule can also bind an α-helix with high affinity. In many of the RNA binding sites for arginine-rich motif family members, the peptides also recognize mostly double helical RNA punctuated by small bulges or ordered loops that maintain continuous helical stacking[21]. The RRE IIB RNA binding site for Rev is like this, with two non-Watson-Crick basepairs forming essential recognition elements within the helix, just as R28 recognizes two non-Watson Crick basepairs within a helix[25]. However, the RNA binding sites all have a significantly greater amount of surface area available to contact the peptide because of the deep major groove of the A-form helix and the presence of non Watson•Crick features that are more irregular than a G•T basepair. R28 would be expected to assume a B-form helical conformation, typical of DNA under most conditions, and our CD data suggest little if any deviation from the regular B-form when this molecule is bound to Rev. The restriction of essential amino acids to ~one-third of the surface of the α-helix is also indicative of a shallow binding site like the B-form major groove.

Therefore, a relevant comparison for the Rev-R28 interaction may be to the many dsDNA binding proteins that use α-helices to recognize the major groove of their target sites. However, again Rev is unique in that the affinity it shows as an isolated α-helix for R28 DNA is quite high. We measure a dissociation constant of 2.4 nM at high salt (210mM monovalent cations) and we observe a strong dependence of affinity on cation concentration, consistent with an affinity ~2 orders of magnitude higher under physiological salt conditions. Generally, DNA binding proteins do not bind with high affinity as monomers, but instead use non-covalent dimers or multiple repeats of a single

DNA-binding domain to achieve high affinities and specificities. It has been proposed, however, that depending on monomer levels within the nucleus, some dimeric transcription factors might recognize their target sequences as monomers and dimerize while bound to DNA[56]. Where affinities of these monomers for DNA have been measured, they are generally on the order of 50 nM, even under low salt conditions [57]. Even in the best cases, where artificial scaffolds have been designed and optimized to stabilize DNA recognition helices, affinities are still in the low nM range at physiological salt concentrations [58-60].

A couple of characteristics probably account for the relatively high affinity of the Rev monomer for R28 DNA. One is the exclusive use of arginines to recognize R28. The ability of arginine to simultaneously make multiple sequence-specific hydrogen bonds, electrostatic contacts, and van Der Waal's interactions allows a single residue to contribute to the binding affinity in several ways. The large electrostatic component of the R28-Rev binding affinity relative to that of Rev for TGTTC DNA or IIB RNA demonstrates that the positive charges of the arginines are recognized at many positions in the binding site. Additionally, the G•T/CG basestep appears to exploit these possibilities in a way that may be adapted for a pair of arginines.

Finally, it is important to note that Rev does show a definite preference for a bulged residue at the base of the binding site. This may produce a unique backbone conformation or other helical disruption which is contacted by Rev. It is also noteworthy that when gel shift assays are performed at 4°, this mutation is more deleterious, suggesting the bulge may also have a role in introducing flexibility into the DNA.

Implications for ssDNA recognition

How do our results relate to the recognition of ssDNA in a biological setting? The molecules that we selected to serve as examples of how ssDNA tertiary structure would promote high affinity DNA recognition ended up, at least in the case of the R28 sequence, deviating only minimally from the double-stranded form. Of all non Watson-Crick basepairs, G•T is the least destabilizing to the DNA helix and the conserved adjacent CG basepair is the most stable neighbor for the G•T[61]. A one residue bulge is the only other essential feature in R28 not found in dsDNA and, were this hairpin to extrude as a cruciform from chromosomal DNA, the function of the bulge might be assumed by the helix junction. Therefore, the stability, relative to dsDNA, that this structure should have as part of a cruciform is probably as high as could have been expected, raising the probability that it could form in an in vivo context. Furthermore, the two G•Ts and the bulge are three independent markers which clearly identify this structure as ssDNA and will increase specificity by limiting competition for binding from the vast excess of dsDNA in the cell.

Our DNA molecules also have regions of nonessential DNA sequence which could be altered to enhance the stability of a hairpin and/or the kinetics of cruciform formation. For instance, the single-stranded hairpin binding site for the phage N4 RNA polymerase is known to be especially stable because of a GC basepair at the top of the stem and an AAG loop, which is also thought to enhance the kinetics of hairpin extrusion[62,63]. These or other cruciform-promoting sequences could be substituted into the R28 hairpin or the conserved TGTTC helix of the 3-helix family to promote in vivo availability without reducing affinity for the protein.

Therefore, it appears that there is some compatibility between the protein recognition features we have selected and those characteristics which might be necessary for biological function. Based on this and the precedent set by the enkephalin hairpin, we would suggest that the results of our selection could be used to identify chromosomal sites which might function in the single-stranded state to specifically bind proteins. Computational searches have been conducted on prokaryotic genomes for inverted repeats that would form stable cruciforms and other hairpin stabilizing criteria have also been included[62]. Screening these potential hairpins for the presence of G•T/CG basesteps should identify a class of structures that are more likely to form specific protein binding sites. Since we understand the protein features required to recognize this basestep, it will also be easier to verify protein partners which mediate the function of these sites. The simplicity with which these high affinity sites are formed and recognized gives us confidence that they and other ssDNA structures will be used more often by nature than is currently appreciated.

Materials and Methods

Rev-agarose preparation

The Rev peptide TRQARRNRRRRWRERQRAAAARC was synthesized with an N-terminal succinate and a C-terminal amide and HPLC purified. For agarose coupling, 1 ml of packed ω-amino-hexyl agarose (4% agarose, epoxy activated with 12 atom spacer; Sigma) was incubated for 30 min. at 25° with freshly prepared 2.5 mM sulfo-SMCC (Pierce) in 5mL 50mM sodium phosphate, pH 7.4. Activated agarose was washed 3 times with buffer and resuspended in 2mL buffer containing 40μg peptide. This was incubated for 2hrs at 25°, and unreacted resin was blocked with 20 μL of 500mM DTT for 30 min. Resin was washed 3 times and resuspended at 40 μg/mL.

In vitro selection

A DNA library of 50 nucleotides with the sequence 5'CGTACGGTCGACGCTAGC $(N)_{50}$ CACGTGGAGCTCGGATCC was synthesized and purified by PAGE. All rounds of selections were carried out in 10mM Tris pH 7.5, 5mM KCl, 5mM $CaCl_2$, 3mM $MgCl_2$, .005% Triton X-100 with NaCl at 400mM (Rounds 1-5), 500mM (Rounds 6-9), and 600mM (Rounds 10-13). For the initial round, 20μL Rev-agarose was incubated in 100μL with 100 pmol library DNA for 30 min. at 4°, and washed 3 times with 400μL buffer containing 5μg yeast tRNA (Invitrogen). DNA was eluted in two 150μL washes in buffer containing 5μM Rev (succTRQARRNRRRRWRERQRAAAARam) and 5μg yeast tRNA . Under these conditions, ~5% of library DNA was retained and ~35% eluted, compared to values of 28% retention and 93% elution for RRE IIB. Eluted DNA

was pooled, ethanol precipitated with 10μg glycogen carrier and PCR amplified. Amplifications were done with 150μM dNTPs, 500nM primer, in the presence of 3mM $MgCl_2$ with 1 min. melting at 94°, 1min. annealing at 55°, and 1 min. extension at 72°. All amplifications were for 10-15 cycles and used the 5' biotinylated primer GGATCCGAGCTCCACGTG. DNA was purified as described[64]. Recovered ssDNA was quantitated by the ethidium bromide spot assay[65].

For all subsequent rounds, 7-20 pmol DNA and 3μg tRNA were incubated with 15μL resin as above with the following changes: in rounds 6-11, elutions were done with 1μM Rev and rounds 12-13 with 500nM Rev; in rounds 6-13, the number of Rev elutions was reduced to 1; in rounds 6-7, prior to elution, beads were washed once with buffer containing 5μM Rev R->K peptide, where all arginines in Rev are mutated to lysine; in rounds 8-13, two Rev R->K washes we done prior to elution; in rounds 8-11, 10μM Rev R->K was used in the washes and for rounds 12-13, 20μM Rev R->K was used. After 13 rounds DNA was amplified, digested with BamHI and SalI, cloned into pUC19, and sequenced.

Modification interference

DNA molecules were 5' P^{32} end-labeled with T4 polynucleotide kinase and purified twice with a Qiagen nucleotide removal column. For dimethylsulfate (DMS) modification, labeled DNA in modification buffer (50mM sodium cacodylate/1mM EDTA/2μg salmon sperm DNA (Invitrogen)) was modified with 1μL DMS (Sigma) for 2 min. at 25°. The reaction was stopped by adding sodium acetate to .3M,5 μg glycogen, and 3 volumes ethanol. Reactions were twice ethanol precipitated and once precipitated

with 50μL .4M NaCl/ 1mL ethanol. For N-ethyl-N-nitrosourea (ENU) modification, DNA in modification buffer was modified for 90 sec. at 80° with 100μL of a freshly made saturated solution of ENU (Sigma) in ethanol. The reaction was stopped as above and twice ethanol precipitated. For hydrazine modification, DNA in 20μL of water was modified for with 30μL hydrazine (Sigma) for 4 min. at 25°. The reaction was stopped as above and twice ethanol precipitated.

Modified DNA was resuspended in 50μL water and fractionated using the protocol described for the in vitro selection. 3μL of Rev agarose and 500mM NaCl binding buffer were used and DNA was eluted with 2μM peptide. After fractionation, eluted DNA was ethanol precipitated. For cleavage at DMS and hydrazine modifications, DNA was suspended in 25μL 1M piperidine (Sigma) with 2μg salmon sperm DNA and incubated for 30 min. at 90°. DNA was vacuum dried, resuspended in 25μL water and dried again, ethanol precipitated once and resuspended in formamide. For cleavage at ENU modifications, fractionated DNA was resuspended in 30μL 15mM sodium phosphate pH 7.3. 5μL 1M NaOH was added and reaction was incubated for 30 min. at 90°. The reaction was neutralized with 5μL HCl, twice ethanol precipitated, and resuspended in formamide. All samples were analyzed on 12% TBE/urea/PAGE gels.

Fluorescence-based binding assay

The procedures described in this section are adapted from the work of Luedtke and colleagues[66]. The Rev peptide describe above was fluoresceinated on its C-terminal cysteine by incubating at 50μM with 500μM 5-(iodoacetamido)fluorescein (from a

10x stock in DMF) in 20mM sodium phosphate pH 8.0/2mM EDTA. The reaction was incubated at 25° for 2 hr. in the dark and purified by HPLC. Labeled peptide was detected and quantified by fluorescein absorbance at 475 nm.

For binding assays, fluorescein-labeled Rev (2.5nM) was mixed with DNA in 30mM Hepes pH 7.5/100mM KCl/ 40mM NaCl/ 10mM ammonium acetate/ 10mM guanidinium•HCl/ 2mM $MgCl_2$/ .5mM EDTA/ .001% Nonidet P-40. Binding reactions were 20μL in 384-well plates and were incubated for 30 min. Fluorescence intensities and anisotropies was measured in a LJL Biosystems Criterion fluorimeter with fluorescein filter sets (ex= 485nm, em=530 nm) and a G-factor of .8. All points were measured in quadruplicate and all values are the average of at least three independent experiments. Binding of molecule A3 was measured by monitoring a DNA-dependent increase in peptide anisotropy, but the anisotropy change upon R28 binding was small and difficult to quantitate. However, a DNA-dependent increase in peptide fluorescence intensity was observed that was qualitatively similar to results obtained from anisotropy and gel-shift experiments. Therefore, this was used as the measure of R28 binding. All data were fit to a single-site binding model using Kaleidagraph.

Competition binding assay

Competition assays were performed by incubating competitor peptide with 2.5nM fluoresceinated Rev and 16nM R28 DNA or 10nM fluoresceinated Rev and 32nM A3 DNA. Reactions were equilibrated for 30 min. at 25° prior to assay. Data were fit to a single site competition model and IC_{50} values were determined.

Gel shift assay

Gel shifts were performed in 10mM Hepes pH 7.5/100mM KCl/1mM $MgCl_2$/.5mM EDTA/50μg/mL yeast tRNA/10% glycerol at 4° using ^{32}P-end labeled DNA at a concentration of less than .5nM. Reactions were performed in 10μL volumes and incubated for 30 min., prior to separation on 10% polyacrylamide/.5xTBE gels.

Circular Dichroism: CD spectra were measured using an Aviv model 62DS spectropolarimeter. Samples were prepared in 10 mM sodium phosphate pH 7.5/100mM KCl/40mM NaCl/2mM $MgCl_2$ at 25°. Nucleic acid and peptide concentrations were 2.5μM.

References

1. Dai, X. & Rothman-Denes, L. DNA structure and transcription. *Current Opinion in Microbiology* **2**, 126-130 (1999).

2. Ding, J., Hayashi, M., Zhang, Y., Manche, L., Krainer, AR & Xu, R. Crystal structure of the two-RRM domain of hnRNP A1 (UP1) complexed with single-stranded telomeric DNA. *Genes and Development* **13**, 1102-1115 (1999).

3. Braddock, D., Louis, J., Baber, J., Levens, D. & Clore, G. Structure and dynamics of KH domains from FBP bound to single-stranded DNA. *Nature* **415**, 1051-1056 (2002).

4. Lei, M., Podell, E., Baumann, P. & Cech, T. DNA self-recognition in the structure of Pot1 bound to telomeric single-stranded DNA. *Nature* **26**, 198-203 (2003).

5. Horvath, M., Schweiker, V., Bevilacqua, J., Ruggles, J. & Schultz, S. Crystal structure of the Oxytrichia nova telomere end binding protein complexed with single strand DNA. *Cell* **95**, 963-974 (1998).

6. Mitton-Fry, R., Anderson, E., TR, H., Lundblad, V. & Wuttke, D. Conserved structure for single-stranded telomeric DNA recognition. *Science* **296**, 144-147 (2002).

7. Mitton-Fry, R., Anderson, E., Theobald, D., Glustrom, L. & Wuttke, D. Structural basis for telomeric single-stranded DNA recognition by yeast cdc13. *Journal of Molecular Biology* **338**, 241-255 (2004).

8. Woodson, S. & NB, L. Structure and dynamics of ribosomal RNA. *Current Opinion in Structural Biology* **8**, 294-300 (1998).

9. Cate, J. *et al.* RNA tertiary structure mediated by adenosine platforms. *Science* **273**, 1696-1699 (1996).

10. Wu, H. *et al.* A novel family of RNA tetraloop structure forms the recognition site for Saccharomyces cerevisiae RNase III. *The EMBO Journal* **20**, 7240-7249 (2001).

11. Tishchenko, S. *et al.* Detailed analysis of RNA-protein interactions within the ribosomal protein S8-rRNA complex from the Archeon Methanococcus jannaschii. *Journal of Molecular Biology* **311**, 311-324 (2001).

12. Gutell, R., Cannone, J., Shang, Z., Du, Y. & Serra, M. A story: Unpaired adenosine bases in ribosomal RNAs. *Journal of Molecular Biology* **304**, 335-354 (2000).

13. Correll, C. & Swinger, K. Common and distinctive features of GNRA tetraloops based on a GUAA tetraloop structure at 1.4 angstrom resolution. *RNA* **9**, 355-363 (2003).

14. Glucksmann-Kuis, M.A., Dai, X., Markiewicz, P. & Rothman-Denes, L.B. E. coli SSB activates N4 virion RNA polymerase promoters by stabilizing a DNA hairpin required for promoter recognition. *Cell* **84**, 147-154 (1996).

15. Spiro, C., Richards, J., Chandrasekaran, S., Brennan, R. & McMurray, C. Secondary structure creates mismatched base pairs required for high-affinity binding of cAMP response element-binding protein to the human enkephalin enhancer. *Proceedings of the National Academy of Science* **90**, 4606-4610 (1993).

16. Spiro, C., Bazett-Jones, D., Wu, X. & McMurray, C. DNA structure determines protein binding and transcriptional efficiency of the proenkephalin cAMP-responsive enhancer. *Journal of Biological Chemistry* **270**, 27702-27710 (1995).

17. Spiro, C. & McMurray, C. Switching of DNA secondary structure in proenkephalin transcriptional regulation. *Journal of Biological Chemistry* **272**, 33145-33152 (1997).

18. Chen, S. & L, S. Identification of long intergenic repeat sequences associated with DNA methylation sites in Caulobacter crescentus and other α-proteobacteria. *Journal of Bacteriology* **185**, 4997-5002 (2003).

19. Dott, P., Chuang, C. & Saunders, G. Inverted repetitive sequences in the human genome. *Biochemistry* **15**, 4120-4125 (1976).

20. Tomonaga, T. & Levens, D. Heterogeneous Nuclear Ribonucleoprotein K is a DNA-binding transactivator. *Journal of Biological Chemistry* **270**, 4875-4881 (1995).

21. Weiss, M. & Narayana, N. RNA recognition by arginine-rich peptide motifs. *Biopolymers* **48**, 167-180 (1999).

22. Faber, C., Scharpf, M., Becker, T., Sticht, H. & Rosch, P. The structure of the coliphage HK022 Nun protein-λ-phage boxB. *Journal of Biological Chemistry* **276**, 32064-32070 (2001).

23. Legault, P., Li, J., Mogridge, J., Kay, L.E. & Greenblatt, J. NMR structure of the bacteriophage lambda N peptide/boxB RNA complex: recognition of a GNRA fold by an arginine-rich motif. *Cell* **93**, 289-299 (1998).

24. Ye, X. *et al.* RNA architecture dictates the conformations of a bound peptide. *Chem. Biol.* **6**, 657-669 (1999).

25. Ye, X., Gorin, A., Ellington, A.D. & Patel, D.J. Deep penetration of an alpha-helix into a widened RNA major groove in the HIV-1 rev peptide-RNA aptamer complex. *Nat. Struct. Biol.* **3**, 1026-1033 (1996).

26. Ye, X., Kumar, R.A. & Patel, D.J. Molecular recognition in the bovine immunodeficiency virus Tat peptide-TAR RNA complex. *Chem. Biol.* **2**, 827-840 (1995).

27. Puglisi, J.D., Tan, R., Calnan, B.J., Frankel, A.D. & Williamson, J.R. Conformation of the TAR RNA-arginine complex by NMR spectroscopy. *Science* **257**, 76-80 (1992).

28. Puglisi, J.D., Chen, L., Blanchard, S. & Frankel, A.D. Solution structure of a bovine immunodeficiency virus Tat-TAR peptide-RNA complex. *Science* **270**, 1200-1203 (1995).

29. Battiste, J.L. *et al.* Alpha helix major groove recognition in an HIV-1 Rev peptide-RRE RNA complex. *Science* **273**, 1547-1551 (1996).

30. Jiang, F. *et al.* Anchoring an extended HTLV-1 Rex peptide within an RNA major groove containing junctional base triples. *Structure* **7**, 1461-1472 (1999).

31. Cai, Z. *et al.* Solution structure of P22 transcriptional antitermination N peptide-box B RNA complex. *Nat. Struct. Biol.* **5**, 203-212 (1998).

32. Suzuki, M. & Yagi, N. An in-the-groove view of DNA structures in complexes with proteins. *JMB* **255**, 677-687 (1996).

33. Tan, R., Chen, L., Buettner, J.A., Hudson, D. & Frankel, A.D. RNA recognition by an isolated α helix. *Cell* **73**, 1031-1040 (1993).

34. Brodsky, A.S. & Williamson, J.R. Solution structure of the HIV-2 TAR-argininamide complex. *J. Mol. Biol.* **267**, 624-639 (1997).

35. Calnan, B.J., Tidor, B., Biancalana, S., Hudson, D. & Frankel, A.D. Arginine-mediated RNA recognition: the arginine fork. *Science* **252**, 1167-1171 (1991).

36. Baskerville, S., Zapp, M. & Ellington, A. Anti-Rex aptamers as mimics of the Rex-binding element. *Journal of Virology* **73**, 4962-4971 (1999).

37. Ellington, A., Leclerc, F. & Cedergren, R. An RNA groove. *Nature Structural Biology* **3**, 981-984 (1996).

38. Varani, G. & McClain, W. The G•U wobble base pair. *EMBO reports* **1**, 18-23 (2000).

39. Cheng, A., Chen, W., Fuhrmann, C. & Frankel, A. Recognition of nucleic acid bases and base-pairs by hydrogen bonding to amino acid side-chains. *JMB* **327**, 781-796 (2003).

40. Nadassy, K., Wodak, S. & Janin, J. Structural features of protein-nucleic acid recognition sites. *Biochemistry* **38**, 1999-2017 (1999).

41. Schumacher, M., Goodman, R. & Brennan, R. The structure of a CREB bZIP•somatostatin CRE complex reveals the basis for selective dimerization and divalent cation-enhanced DNA binding. *Journal of Biological Chemistry* **275**, 35242-35247 (2000).

42. Isaacs, R., Rayens, W. & Speilmann, H. Structural differences in the NOE-derived structure of G-T mismatched DNA relative to normal DNA are correlated with differences in 13C relaxation-based internal dynamics. *Journal of Molecular Biology* **319**, 191-207 (2002).

43. Keller, W., Konig, P. & Richmond, T. Crystal structure of a bZIP/DNA complex at 2.2 angstroms: determinants of DNA specific recognition. *Journal of Molecular Biology* **254**, 657-667 (1995).

44. Konig, P. & Richmond, T. The X-ray structure of the GCN4-bZIP bound to ATF/CREB site DNA shows the complex depends on DNA flexibility. *Journal of Molecular Biology* **233**, 139-154 (1993).

45. Agalarov, S., Prasad, G., Funke, P., Stout, C. & Williamson, J. Structure of the S15,S6,S18-rRNA complex: assembly of the 30S ribosome central domain. *Science* **288**, 107-112 (2000).

46. Nikulin, A. *et al.* Crystal structure of the S15-rRNA complex. *Nature Structural Biology* **7**, 273-277 (2000).

47. Declais, A.-C. *et al.* The complex between a four-way DNA junction and T7 endonuclease I. *The EMBO Journal* **22**, 1398-1409 (2003).

48. Lilley, D. Structures of helical junctions in nucleic acids. *Quarterly Reviews of Biophysics* **33**, 109-159 (2000).

49. Roe, S. *et al.* Crystal structure of an octameric RuvA-Holliday junction complex. *Molecular Cell* **2**, 361-372 (1998).

50. Ariyoshi, M., Nishino, T., Iwasaki, H., Shinagawa, H. & Morikawa, K. Crystal structure of the Holliday junction DNA in complex with a single RuvA tetramer. *Proceedings of the National Academy of Sciences* **97**, 8257-8262 (2000).

51. Morellet, N. *et al.* Structure of the complex between the HIV-1 nucleocapsid protein NCp7 and the single-stranded pentanucleotide d(ACGCC). *Journal of Molecular Biology* **283**, 419-434 (1998).

52. Raghunathan, S., Kozlov, A., Lohman, T. & Waksman, G. Structure of the DNA binding domain of E.coli SSB bound to ssDNA. *Nature Structural Biology* **7**, 648-652 (2000).

53. Su, L. *et al.* An RNA enhancer in a phage transcriptional antitermination complex functions as a structural switch. *Genes Dev.* **11**, 2214-2226 (1997).

54. Escalante, C., Yie, J., Thanos, D. & Aggarwal, A. Structure of IRF-1 with bound DNA reveals determinants of interferon regulation. *Nature* **391**, 103-106 (1998).

55. Shakked, Z. *et al.* Determinants of repressor/operator recognition from the structure of the trp operator binding site. *Nature* **368**, 469-473 (1994).

56. Park, C., Campbell, J. & Goddard, W. Can the monomer of the leucine zipper proteins recognize the dimer binding site without dimerization? *Journal of the American Chemical Society* **118**, 4235-4239 (1996).

57. Cranz, S., Berger, C., Baici, A., Jelesarov, I. & Bosshard, H. Monomeric and dimeric bZIP transcription factor GCN4 bind at the same rate to their target DNA site. *Biochemistry* **43**, 718-727 (2004).

58. Zondlo, N. & Schepartz, A. Highly specific DNA recognition by a designed miniature protein. *Journal of the American Chemical Society* **121**, 6938-6939 (1999).

59. Montclare, J. & Schepartz, A. Miniature homeodomains: high specificity without an N-terminal arm. *Journal of the American Chemical Society* **125**, 3416-3417 (2003).

60. Chin, J. & Schepartz, A. Concerted evolution of structure and function in a miniature protein. *Journal of the American Chemical Society* **123**, 2929-2930 (2001).

61. Allawi, H. & SantaLucia, J. Thermodynamics and NMR of internal G•T mismatches in DNA. *Biochemistry* **36**, 10581-10594 (1997).

62. Dai, X., Greizerstein, M., Nadas-Chinni, K. & Rothman-Denes, L. Supercoil-induced extrusion of a regulatory DNA hairpin. *Proceedings of the National Academy of Sciences* **94**, 2174-2179 (1997).

63. Dai, X., Kloster, M. & Rothman-Denes, L. Sequence-dependent extrusion of a small DNA hairpin at the N4 virion RNA polymerase promoters. *Journal of Molecular Biology* **283**, 43-58 (1998).

64. Harada, K. & Frankel, A. Identification of two novel arginine binding DNAs. *The EMBO Journal* **14**, 5798-5811 (1995).

65. Moore, D., Chory, J. & Ribaudo, R. *Rapid estimation of DNA concentration by ethidium bromide dot quantitation*, (, 1994).

66. Luedtke, N., Liu, Q. & Tor, Y. RNA-ligand interactions: affinity and specificity of aminoglycoside dimers and acridine conjugates to the HIV-1 Rev response element. *Biochemistry* **42**, 11391-11403 (2003).

67. Bartel, D.P., Zapp, M.L., Green, M.R. & Szostak, J.W. HIV-1 Rev regulation involves recognition of non-Watson-Crick base pairs in viral RNA. *Cell* **67**, 529-536 (1991).

Figure Legends

Figure 1. Selection of single-stranded DNA molecules specific for HIV-Rev. (A) The Rev peptide used for selection and its natural high affinity RNA site, IIB. Bold peptide residues are essential for RNA binding. The succinyl group at the N-terminus and the alanine tail at the C-terminus are added to stabilize the α-helical conformation. Bold RNA residues were conserved in an in vitro selection to identify essential positions[67]. Box indicates a requirement for a purine-purine basepair. (B) Sequences selected to bind Rev. Randomized regions are in capitals, priming sequence is in lowercase. Red indicates conserved TGTTC motif, blue is the AGCA complement. Molecules in which there is TGTTC point mutation that covaries with a point mutation in the AGCA are shown in the lower panel. (C) Summary of TGTTC variants in the selected pool. (D) Basepairing of TGTTC motif and summary of putative 3-helix junction fold. Numbers indicate the range of observed linker sizes between helices.

Figure 2. Rev recognizes TGTTC as part of a 3-helix junction. (A) Representative binding curves derived from measuring the anisotropy of fluorescein-labeled Rev. Shown are two selected DNA molecules, A3 (—▲—), and 29 (—▼—), as well as unselected pool DNA (—◆—), RRE IIB RNA (—●—), and BIV TAR RNA (—■—), a non-specific 28-nucleotide RNA hairpin. (B) Summary of relative dissociation constants for sequence A3 mutants. Values are averages of at least three independent experiments. (C) Representative modification interference results. Modified, unbound DNA, U, is on the left and bound DNA, B, is on the right of each gel. Large dots indicate positions where modification strongly interferes with binding, small

dots indicate moderate interference. (D) Summary of results from (C) Colored positions in capitals represent interference by DMS modification (red) or hydrazine modification (blue). Bold indicates strong interference, plain indicates moderate interference. Large arrowheads indicate strong interference by ENU modification, small arrowheads indicate moderate interference. Data are derived from at least three independent experiments.

Figure 3. Rev also recognizes a mostly double-stranded DNA hairpin. (A) Representative binding curves derived from fluorescence intensity binding assay for R28 DNA and selected mutants. R28 (—●—), T32deletion (—■—), C33T/G50A (—⊞—), G34A/C49T (—+—), G48A (—▲—), G37A (—▼—), C38T/G45A (—◆—). (B) Summary of relative dissociation constants for all R28 mutants. Values are averages of at least four independent experiments. Box indicates basepairs that are conserved in one additional selected sequence. (C) Representative modification interference results. Modified, unbound DNA, U, is on the left and bound DNA, B, is on the right of each gel. Positions where modification significantly affects binding are indicated. (D) Summary of results from (C). Colored positions in capitals represent interference by DMS modification (red) or hydrazine modification (blue). Large arrowheads indicate strong interference by ENU modification, small arrowheads indicate moderate interference. Data are derived from at least three independent experiments.

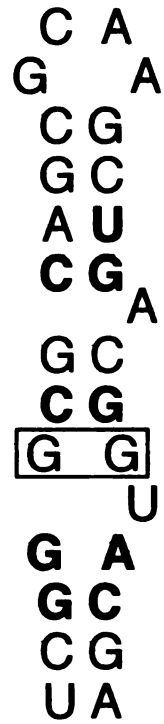
Figure4. R28 DNA adopts a B-form DNA helix. CD spectra of R28 DNA (●), R28 DNA with G•T->AT mutations (○), R28 RNA (◆), and R28 G•T->AT (◇). (A) is free DNA while (B) is in the presence of Rev.

Figure 5. Amino acid requirements for ssDNA recognition. (A) Relative affinities of alanine substitution mutants for R28 and 3-helix DNA. (B) Helical wheel depiction of the Rev α-helix. Positions where alanine substitution reduces affinity by at least 10-fold are circled in red.

Figure 6. Two arginine residues determine specificity for each G•T/CG basestep. (A) Representative gel shifts for wild-type R28DNA with wild-type Rev, as well as the indicated mutant with Rev R38A. For R28DNA with wild-type Rev, peptide concentrations are, from left to right, 1-256nM. For all assays with the R38A peptide, peptide concentrations are 16-8192nM. (B) Binding curves for arginine mutants against R28 DNA mutants. The DNA molecules are wild-type R28 DNA (—▲—), G34A/C49T (—○—),G48A (—●—),G37A (—■—), and C38T/G45A (—□—). Mutants are the same as shown in Figure 3.

Figure 7. Possible schemes for arginine recognition of a G•T/CG basestep. (A) Schematic of the G•T/CG basestep. Hydrogen bond acceptors (thymine O4, guanosine O6 and N7) are in red. (B) Predicted hydrogen bonding arrangement between arginine and a G•T basepair.

Figure 1A



$_{34}$ succ**TR**QA**RRN**RRR**R**W**R**E**R**Q**R**$_{50}$AAAAR

Figure 1B

Sequence	
A6	ctagcACAACGGTGTTCATGAGTACTCGGATCTCTCAGCGGCGACTCGCTCCACGcac
D6	ctagcATACCAATGTTCAGACGGGGATGGGCTGCAGCATCCCCGACTCCGTCGCTcac
G5	ctagcACATCCGTGTTCACGCCAGTCCGTCGAGCTTACGGCGTTGGTCCCCGAGCCAc
D5	ctagcACCTACGGATGTTCATTACTATATTTTGCAAGACTTGCCGCTCCACCAGTcac
F6	ctagcAGCAAATCCTGTTCAAGCGACCATAATCGCGCTCGACGACCGGAGCTTCAcac
33	ctagcACCGCAGGCTGTTCGGCATAACACCGTGTATGACCCGGCGACCGGAGCTCcac
11	ctagcCCAGCATCTTACATGTTCAGGTACCAGACACCTGGTCCTCGCTCCGTCGCcac
31	ctagcGGCAACGTTCTTGTGTTCCCTACGGCTGGACACTCAGCTCGACCTCGCTCCcac
B4	ctagcACAGCTTTTAACTCGTGTTCCCGTCCATGGCGCAGGCGGCCGTTCCAATCcac
67	ctagcAGACCCGAAGGGGAATGTTCAACGTATATGCTTCGTTCCGGAGACGCGTTcac
44	ctagcACATGCGCAGGCCATGTGTTCGGACCCTTCCCGTCGTCCGTGCCCCCATAcac
17	ctagcAGGCTGTTATAAGCTACTGTTCCCGATTTCGCAGGAGAGCGGCCCGAGGTcac
A8	ctagcACAAGACAACTCAGCATACAATGTTCATGGTTTTAAACCTCAGTATTTCAcac
A9	ctagcATGATGTGGTAACGACAGCAATGTTCCTTGCTAGGATAAGCAACAGTCACcac
A12	ctagcCCAGCGACTACTGATAGTAGTTGTTCAACATAATTCCGCATAAtTCAGATcac
E5	ctagcATAGTCAACGGGGCGCGCAAATGTTCCAATCTCGTTTAGCGGCTCCACTAcac
E6	ctagcATAAGCCTGGACGACCTGGATAATGTTCCTACACACTGTGTACAGTCGCAcac
F7	ctagcGGGAGGCAGCGTTATCAAAGAGAATGTTCATCGCACAATGGTGTGCCCCTcac
C7	ctagcACACAGCAAAGTCATTAGACGCCACTAATCTTTGTTCCTTTTATCTAAATcac
D15	ctagcCAGCGGACGGAGGCTCTGAGTGACCCAGCACTTTGTGTTCCGTGCTGCCGcac
29	ctagcACAAAATAGGCCAAGCGGAGCGCAGCACCTTACGGTGTTCCCCCCTATTAcac
55	ctagcACCCTGCTTCAGTGGATCTGTAGTGCTGGCCAAGATGTTCGCTCCACCCCcac
F2	ctagcGGACCAGATCTGACGACCCCCCAGCATCCTTTTTGGGAATGTTCCGTATAcac
39	ctagcCCATGCACGGGGATAGCCCAGCAAAACCCTCTCGCGTTTTGTTCTCGACAcac
E8	ctagcCGAACCGCCCAGCGAATGTCACTTAACCTCTACCGGAGGCAACATTTGTTCac
27	ctagcTAGTGTCGGAAAAACCCCCAGCGGATGTCCTCGGGTAGATGACGTCTGTTCac
40	ctagcACTGCATCCAGGCGCCGTCTCAGCGAAAAGGGTGCAACCTGATTTTTGTTCac
	TGTTC variants with covariation at position 1
B1	ctagcTGCAAGTTCTTCATGTTTTCGAACATGTCAAGTCGCTCCACAGCCACAATcac
E3	ctagcCCCAGGTTCATAATCAAGAGCATTGTTATTCGTCGCTTCCCTCGCTCCACcac
31	ctagcGGCAACGTTCTTGTGTTCCCTACGGCTGGACACTCAGCTCGACCTCGCTCCcac
major	ctagcATGAGCGTAGACACGCATCCCCAGCGAGTTGCAGACTCGTTCAGCTCCAcac
A3	ctagcTAGTGAGGTCGTAGCGTAAATAGGATTCCCCCAGCTGTCTAGTTCAGCTCcac

Figure 1

C

	T	G	T	T	C
Wild type	----------27----------				
Variants	11	4	5	9	4
Covariants	5	0	0	2	x

Representation of each sequence from 50 independent clones (2250 pentuplets)

D

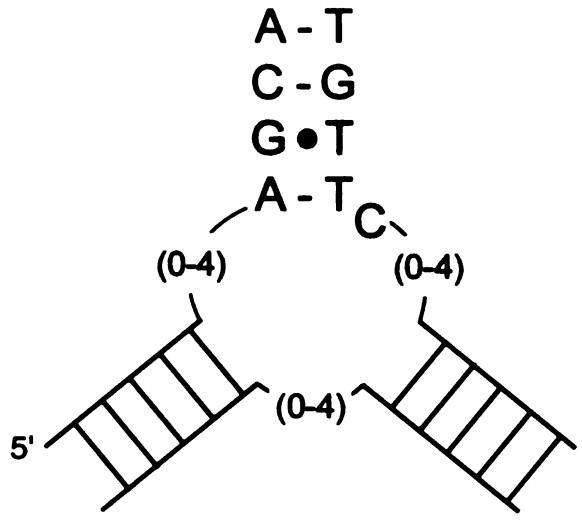
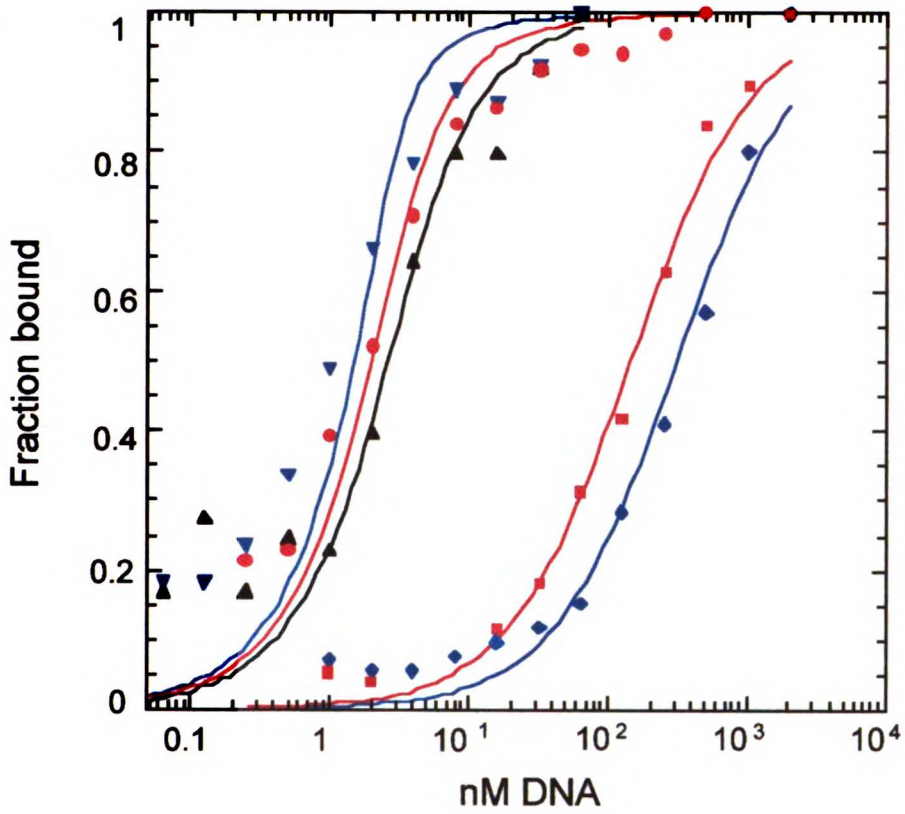


Figure 2

A


Fraction bound
nM DNA
0
0.2
0.4
0.6
0.8
1
0.1
1
10 1
10 2
10 3
10 4

B

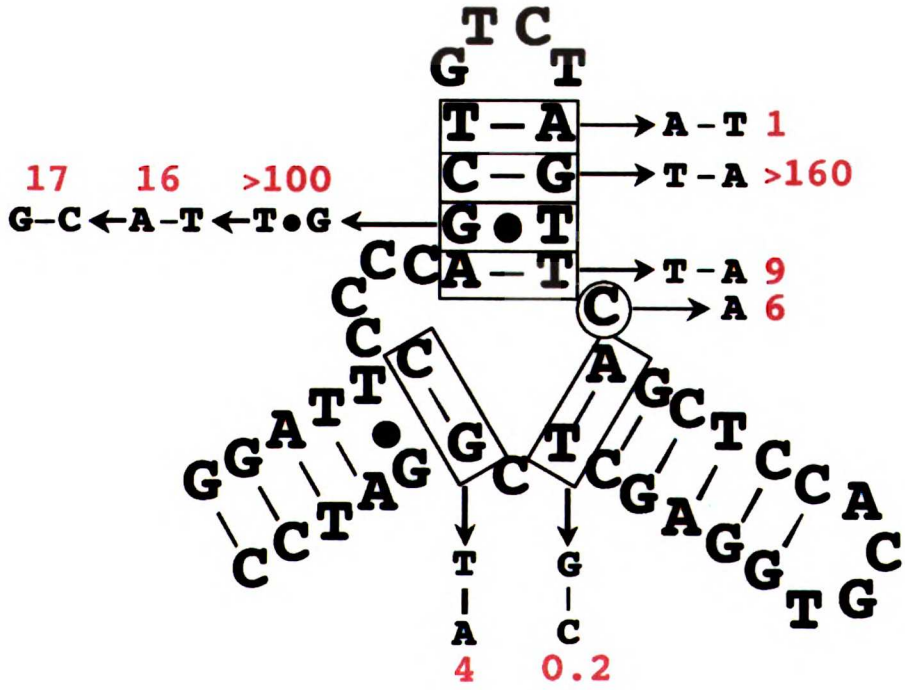
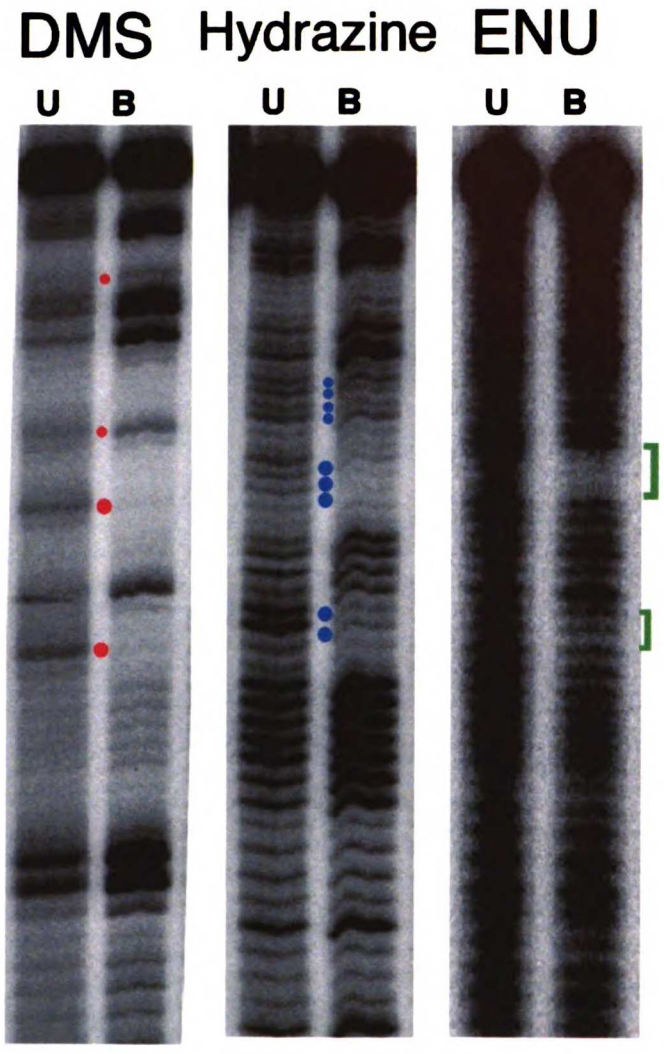

17
16
>100
G–C
A–T
T•G
A–T 1
T–A >160
T–A 9
A 6
4
0.2

Figure 2

C



D

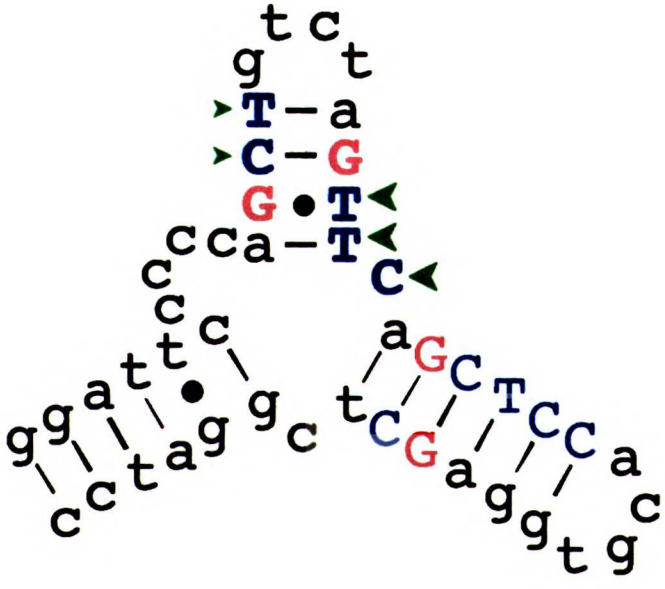
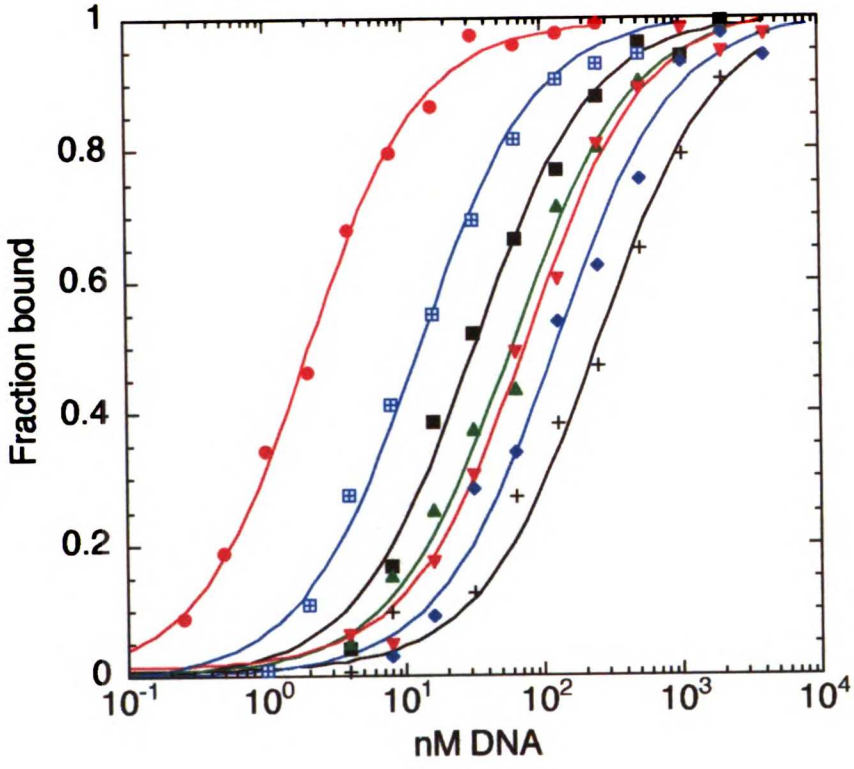


Figure 3

A


Fraction bound
nM DNA
0
0.2
0.4
0.6
0.8
1
10^{-1}
10^{0}
10^{1}
10^{2}
10^{3}
10^{4}

B

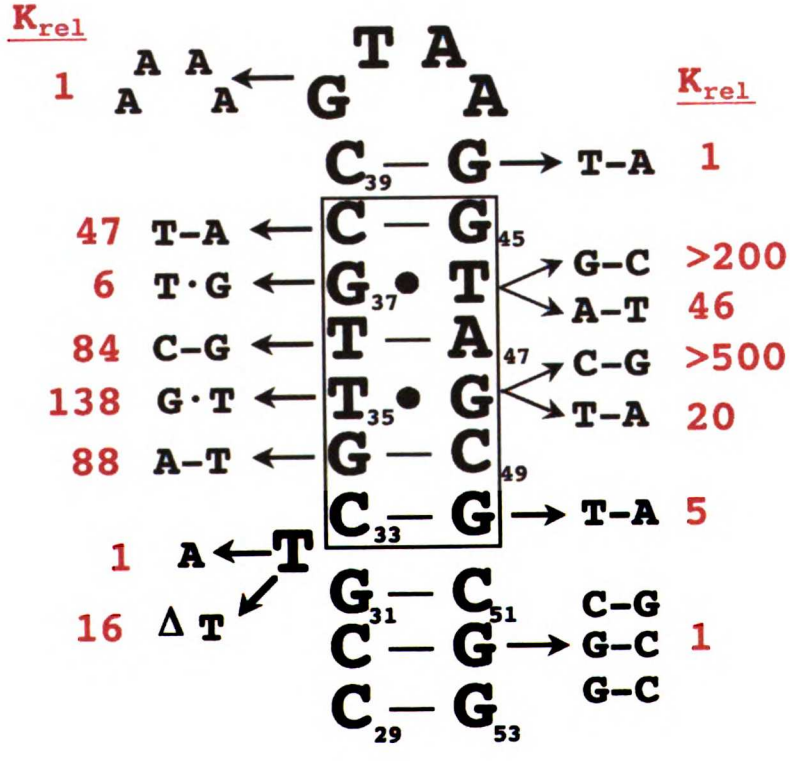
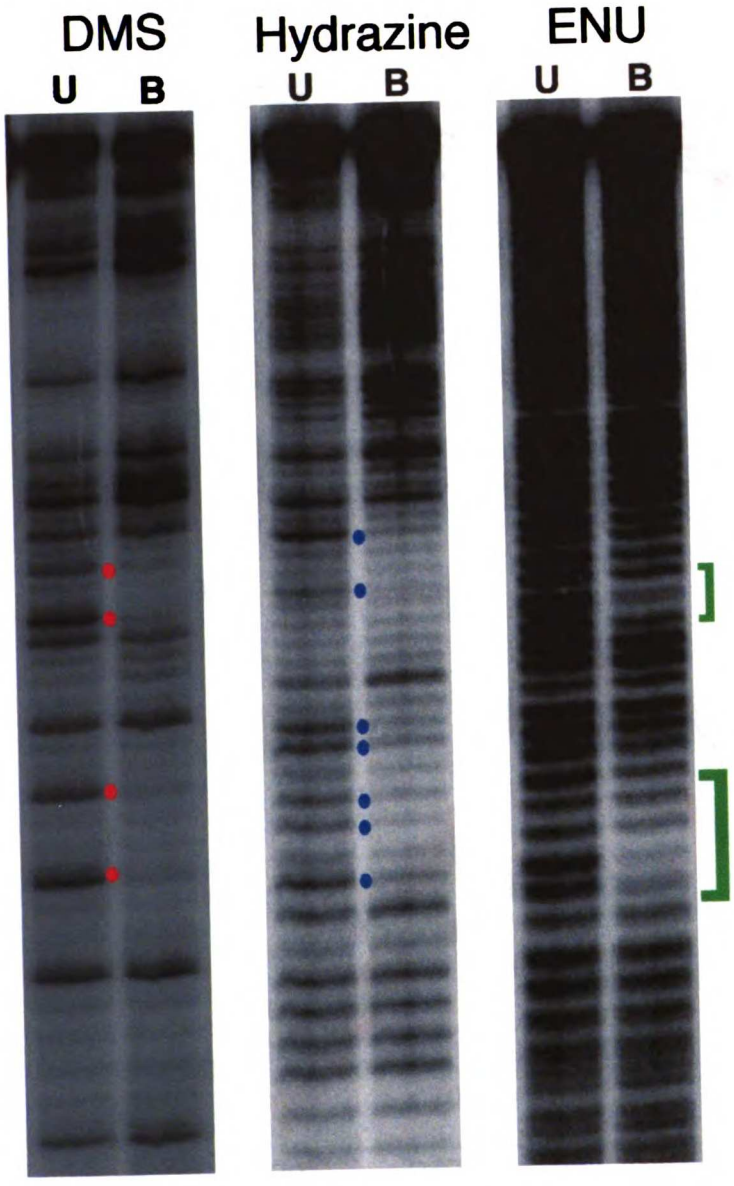

K_{rel}
1
A A
A A
T A
G A
K_{rel}
C39 — G → T-A 1
47 T-A ← C — G45
6 T·G ← G37 • T → G-C >200
A-T 46
84 C-G ← T — A47 → C-G >500
138 G·T ← T35 • G → T-A 20
88 A-T ← G — C49
C33 — G → T-A 5
1 A ← T
16 Δ T
G31 — C51
C — G → C-G G-C 1
G-C
C29 — G53

Figure 3

C


DMS
U B
Hydrazine
U B
ENU
U B

D

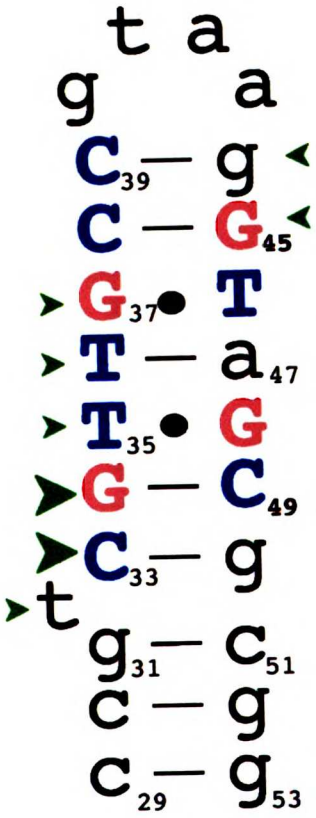

t a
g a
C39 — g
C — G45
G37 • T
T — a47
T35 • G
G — C49
C33 — g
t
g31 — c51
c — g
c29 — g53

Figure 4

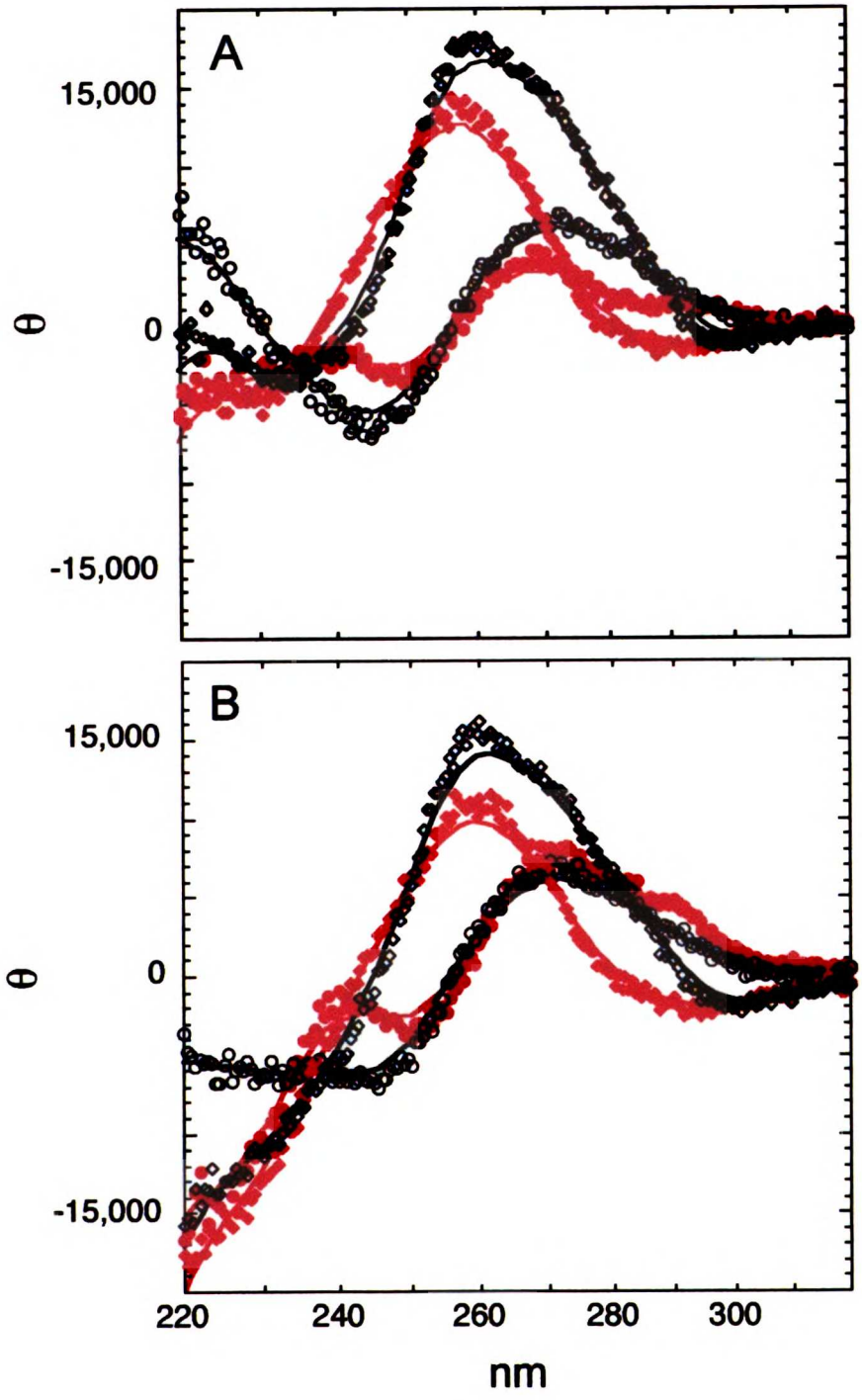
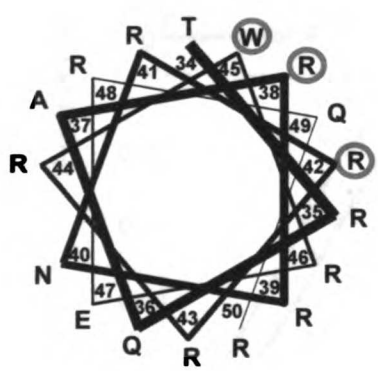


Figure 5

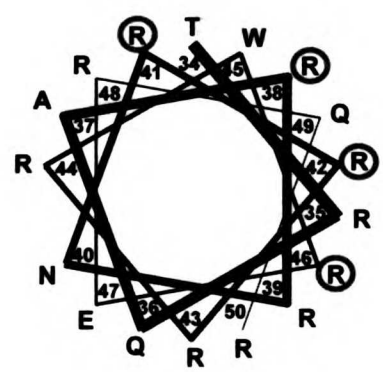
A

	T_{34}	R_{35}	Q_{36}	A_{37}	R_{38}	R_{39}	N_{40}	R_{41}	R_{42}	R_{43}	R_{44}	W_{45}	R_{46}	E_{47}	R_{48}	Q_{49}	R_{50}
3-helix	1	1	1	1	>200	1	1	2	200	2	1	>200	1	1	4	1	2
R28	1	1	1	2	19	2	1	170	9	1	3	4	44	1	2	1	1

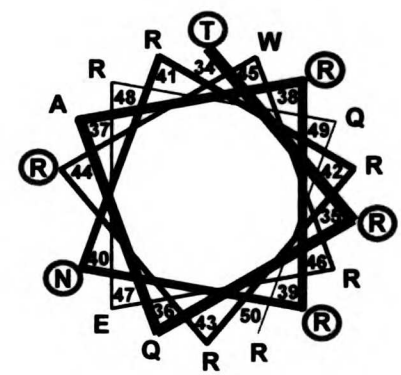
B



3-helix



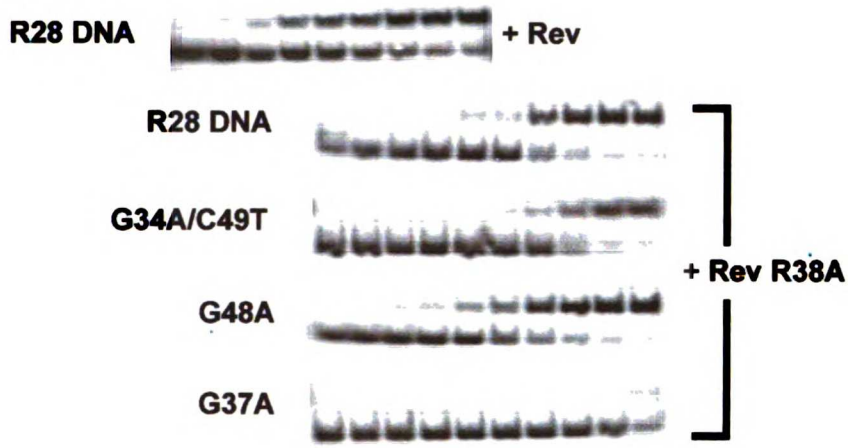
R28



IIB

Figure 6

A


R28 DNA
+ Rev
R28 DNA
G34A/C49T
G48A
G37A
+ Rev R38A

B

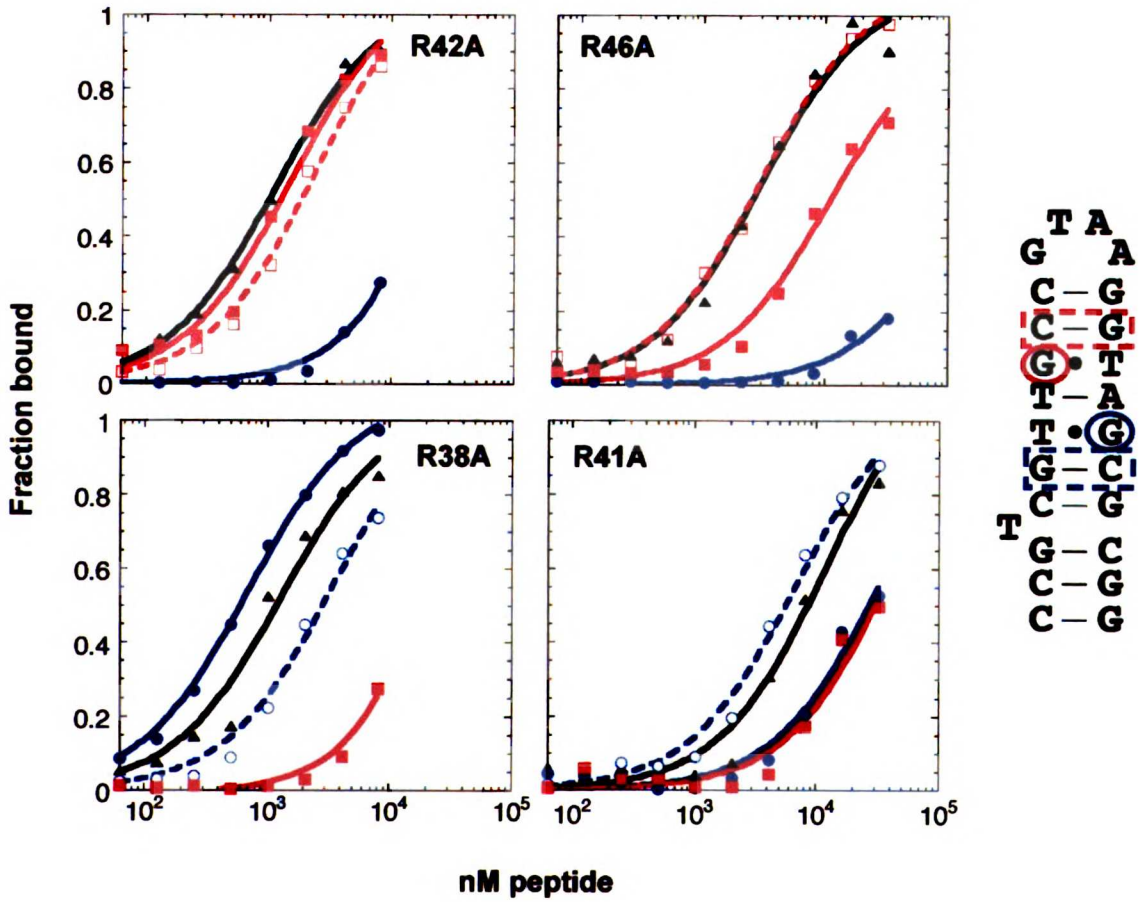
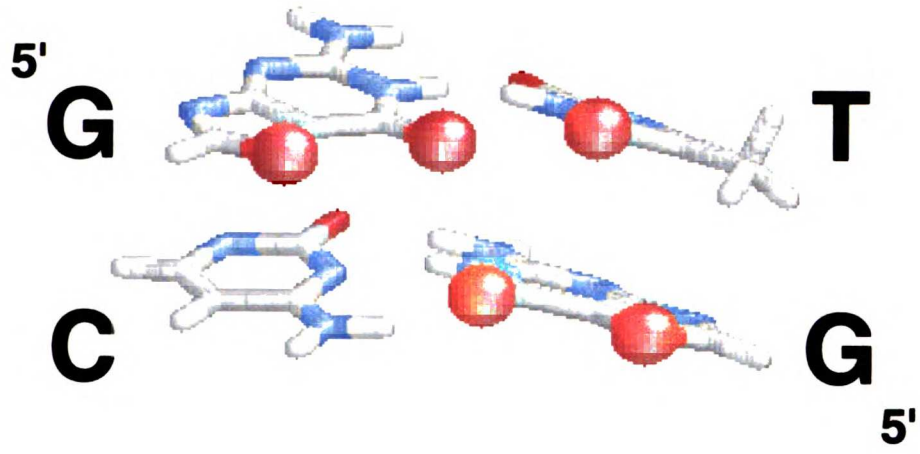
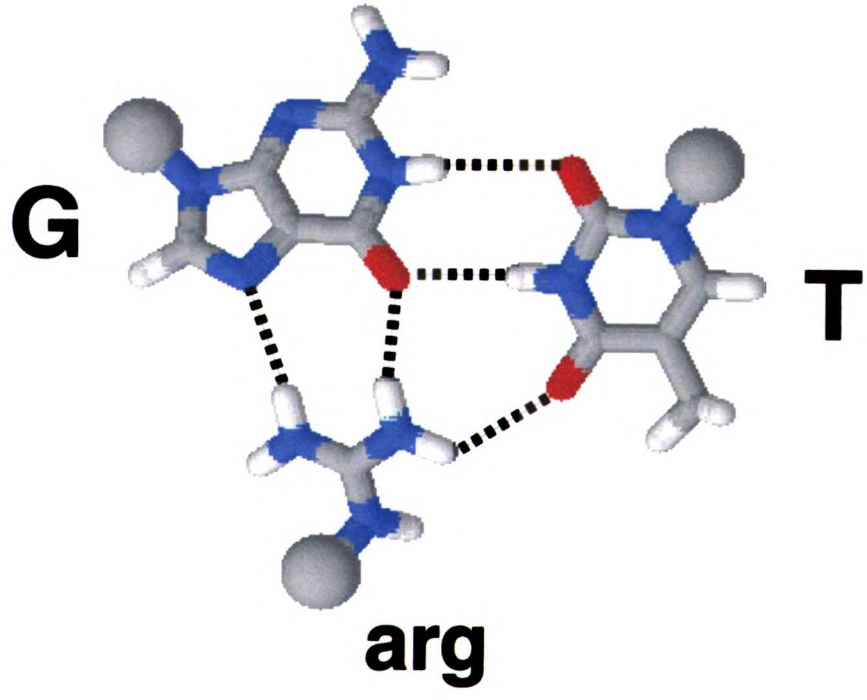

R42A
R46A
R38A
R41A
Fraction bound
nM peptide
1
0.8
0.6
0.4
0.2
0
10^2
10^3
10^4
10^5
T A
G A
C — G
C — G
G • T
T — A
T • G
G — C
C — G
T
G — C
C — G
C — G

Figure 7

A.



B.



Chapter 4

Recognition of branched DNA structures by an arginine-rich peptide

Recognition of branched DNA structures by an arginine-rich peptide

Stephen G. Landt, Alejandro Ramirez, and Alan D. Frankel*

[1]Department of Biochemistry and Biophysics
University of California, San Francisco
San Francisco, CA 94143-2280

*Address correspondence to: Alan Frankel
Department of Biochemistry and Biophysics
UCSF
600 16th Street
San Francisco, CA 94143-2280

Telephone: 415-476-9994
FAX: 415-514-4112
e-mail: frankel@cgl.ucsf.edu

Abstract

In an effort to understand the role nucleic acid tertiary structure can play in protein recognition of single-stranded DNA, we have previously reported the selection of a family of ssDNA molecules that bind to the arginine-rich peptide from HIV Rev with an conserved TGTT/AGCA duplex in the context of a predicted 3-helix junction. To understand the mechanism of branched helix recognition by this small peptide, we mutagenized the predicted 3-helix structure and biochemically characterized the role of an essential tryptophan residue in recognition. We show that Rev binding requires at least three helices and that 3-helix junctions are bound with moderately higher affinities and specificities than cruciforms. Mutagenesis shows that the primary peptide determinant of specificity is a single aromatic amino acid. We also observe that the fluorescence emission of the tryptophan is dramatically quenched in the presence of DNA. Based on this and geometric considerations, we propose that structure specificity derives from the stacking of tryptophan on the terminal basepair of one of the three helices at the junction. This demonstrates a unique mechanism of coupling the recognition of nucleic acid junctions to sequence specific recognition that might be a general strategy for recognizing ssDNA.

Introduction

Branched helical structures are involved in almost all aspects of nucleic acid function[1]. In DNA metabolism, 4-way Holliday type junctions are the necessary intermediates for recombination and repair. In large RNA structures, the intersection of multiple helices, especially 3-helix junctions, are critical for organizing the folding of large domains and in establishing catalytically active structures [2,3]. Even DNA 3-helix junctions have been documented in specific locations, such as ssDNA viral genomes and in certain promoter derangements associated with triplet expansion diseases [4-7].

3-helix junctions have also been shown, at least for RNA, to act as specific binding sites for proteins and understanding how these structures are recognized will be an important step in understanding their functions. The best characterized of these interactions is between the central domain of 16S rRNA and the ribosomal protein S15[8-11]. In this case, S15 interacts almost exclusively with the minor groove of one of the three helices, but it also recognizes the branched structure with contacts to the phosphate backbone at the junction itself, which assumes a unique conformation to accommodate the helix intersection[12]. This exemplifies one mechanism by which a protein can recognize a branched structure, a mechanism that is particularly suited to RNA, where the shape of the A-form helix limits accessibility to the bases themselves. However, there are certainly other ways by which branched nucleic acids may be recognized which should be more applicable to other RNA or DNA junctions.

We recently carried out an in vitro selection using the arginine-rich domain from the RNA-binding protein HIV Rev as a target for ssDNA ligands (in preparation). There are

several well-characterized examples in the literature of DNA structures that are recognized specifically in the single-stranded form by proteins as part of regulatory processes, as well as other, less characterized examples which suggest this phenomenon may not be uncommon [13-22]. Based on lessons learned from the study of protein:RNA recognition, it is likely that nucleic acid secondary and tertiary structure will usually be important for facilitating tight, specific binding of these sorts of sites. The goal of our selection, therefore, was to use a nucleic acid binding motif adapted to structured RNA as a means to identify model ssDNA ligands that are recognized, at least in part, by their tertiary structure. The dominant outcome from the selection was a family of sequences that contained the specific sequence motif TGTTC as part of one helix in an apparent 3-helix junction structure. We have shown that the essential sequence feature of this and other families of selected molecules is a G•T/CG basestep that serves as a receptor for a pair of arginine sidechains. Now, it is our goal to determine the extent to which tertiary structure can determine specificity for ssDNA and to identify the mechanisms by which the DNA 3-helix junction is recognized.

We used a set of mutations in the branched structure to identify and measure the contributions of different features of the 3-helix junction to binding affinity and specificity. We also used a series of peptide mutants to dissect the role of a tryptophan residue that is essential for establishing structure specificity. Together, these results show that the Rev peptide derives at least as much specificity from the 3-helix structure as it does from the selected sequence elements and that most of this is probably achieved by stacking of the tryptophan side chain on one of the helix termini. This is the first evidence of this type of mechanism being used as a major specificity determinant in a

high affinity protein:DNA complex and suggests a strategy for proteins to recognize specific branched structures as they may occur in cellular ssDNA.

Results

From our selection, we identified 28 molecules that contained the pentameric sequence TGTTC and its partial complement AGCA. When these molecules were computationally folded, the lowest energy prediction for the TGTTC/AGCA helical pairing was as part of a 3-helix junction fold in 22 of these cases, and the folds of several of the remaining molecules predicted three consecutive helices, although not constrained relative to one another [23]. Figure 1 shows the predicted 3-helix fold of two representative sequences, A3 and 29, which bind Rev with affinities near 1nM and have been extensively characterized. These two molecules illustrate the variation apparent within the predicted structures- the linkers between the helices vary between 0 and 4 nucleotides and there are no conserved sequences aside from a weak preference for runs of C residues. When these molecules were subjected to chemical modification interference, it appeared that there were many important contacts with the conserved TGTTC helix, while some additional, less important interactions were being made to residues predicted to be in one of the adjacent helices. This is generally consistent with the predicted 3-helix fold and suggests a configuration in which the lengths of two of the helices are in relative proximity to one another. However, other structural models are not excluded by these results.

To obtain direct and quantitative evidence that a 3-helix junction is the functional structure, we made a series of truncations in both the A3 and 29 molecules. We separately eliminated each of the helices containing nonessential sequences and linked the remaining helix to the TGTTC/AGCA arm with either a short tether, a longer more

flexible linker derived from the parent sequence, or no linker at all to provide all possible flexibility. In all cases, the computationally predicted structure was a two-helix fold, as shown in Figure 2b [23]. If the essential bound structure is simply two helices adjacent to one another rather than a true 3-helix junction, these less constrained linkages might be expected to allow peptide binding with reasonable affinity. The results of the truncations are shown in Figure 2.

It is clear that all the truncated molecules bind Rev very poorly, at least 100-fold weaker than the parent molecule and generally much weaker than that. By comparison, the most deleterious point mutation that we identified, a CG->TA switch in the TGTTC helix (see Figure 3), reduces binding between 100 and 400-fold. This indicates that the tertiary structure makes a contribution to the binding affinity that is quite comparable to that made by the conserved sequence. When truncations of the same helix but with different linkers are compared, it is usually the case that the flexible linkage allows ~5-fold tighter binding. However, in each of these cases, binding is still reduced by several hundred-fold, demonstrating that a specific feature of a branched helical structure is essential. It is also interesting that truncating the helix suggested by modification interference to directly contact the peptide has an effect that is similar to truncating the non-contacted helix in the case of molecule 29 and actually has a significantly smaller effect than truncating the non-contacted helix in molecule A3. This suggests that contacts to a second helix are nonspecific and that it may be a feature of the junction itself rather than the orientation of a second helix that most determines structure specificity.

To assess the specificity of this interaction for a 3-helix junction, we constructed additional mutants in which a fourth helix was added to the A3 sequence, generating a cruciform structure. The junction of a cruciform should be broadly similar to that of a 3-helix structure- both will present pairs of helices in proximity to one another and have a junction characterized by multiple helix ends and a concentration of negative charges- but the specific conformation at the junction will certainly vary in its details[1]. We introduced the additional helix opposite the conserved TGTTC arm (Figure 3) to preserve the orientation between the selected helices and used either the selected sequences to link the helices or we reduced all linkers to one residue to more naturally mimic a cruciform extruded from dsDNA. As specificity controls, we also introduced the strongly deleterious CG->AT mutation described above into the conserved TGTTC helix of both cruciforms.

Figure 3 shows that both cruciforms bind Rev identically and are far better binders than any of the two-helix structures shown in Figure 2. However, both have an approximate 5-fold reduction in binding affinity relative to the 3-helix junction and a much higher affinity for the TGTTC mutant than when it is present in the 3-helix junction, with an overall reduction in specificity of about 35-fold. It is likely that the large positive charge on the Rev peptide is recognizing the greater negative charge density near the cruciform junction or perhaps a more favorable juxtaposition of helical backbones[24]. In either case, the loss of specificity indicates that the mode of binding to the cruciform is somewhat different than to the 3-helix junction and that Rev is recognizing something specific about the junction of three helices.

One additional result also shows the specificity of Rev for 3-helix junctions. With both the A3 and 29 sequences, peptide binding results in an increase in DNA mobility in native TBE PAGE gels (Figure 4). This is unexpected given the large positive charge on the peptide and that its mass is significant compared to that of the DNA. However, this phenomenon has been reported previously, always associated with 3-helix junction binding sites, and is taken to indicate a protein-induced conformational change, where two of the helices coaxially stack and the third protrudes at an acute angle with respect to the axis of the others [9,25,26]. Although it is unclear whether this helical arrangement contributes to binding or whether it is a side effect of a specific junction conformation, it should place the unstacked helix close enough to one of the stacked arms to account for the identification of contacts to two helices in the modification interference experiments.

As another approach to understanding recognition of these molecules, we tried to characterize the peptide contribution to 3-helix specificity. We have previously shown that only three amino acid residues, arginines 38 and 42 and tryptophan 45, are essential for binding the 3-helix/TGTTC family of molecules and that the essential G•T/CG basestep in the TGTT helix is probably the binding site for the two arginine residues. By process of elimination, the tryptophan residue is therefore likely to be primarily responsible for recognizing the 3-helix structure. Tryptophans have important roles in many specific protein-nucleic acid recognition events and have generally been shown to directly contribute to recognition by stacking on bulged and looped residues in RNA, or by hydrogen bonding to the phosphate backbone in DNA complexes [27-29]. In the context of RNA recognition, however, the tryptophan in Rev does not have an important role[30]. To help determine the role of tryptophan in our case, we made a series of mutants

at position 45 and tested their affinity for the 3-helix molecules (Figure 5). While an alanine substitution was previously shown to dramatically reduce binding, our results show that substitution with phenylalanine or tyrosine allows binding that is indistinguishable from wild-type. Surprisingly, substitution with the nonpolar leucine sidechain has a more deleterious effect than alanine substitution. This identifies the aromaticity of the sidechain as the essential feature for recognition, which supports the notion that the tryptophan is primarily required for its ability to stack on basepairs at the ends of one or more of the helices. It also suggests that any hydrogen bonds or other polar interactions are of little consequence.

To further establish that the tryptophan is involved in base stacking, we analyzed the effects of different nucleic acids on the fluorescence emission spectra of Rev. Several studies show that tryptophan stacking is an important determinant of non-specific affinity for nucleic acids and that nucleic acid binding dramatically quenches tryptophan fluorescence in these complexes[31-34]. When spectra of Rev are taken at low salt concentrations (Figure 6A), we observe ~90% quenching in the presence of either sequence A3 or 29. We also see significant quenching when mutations in each sequence, either the BC1 truncations from Figure 2 or the CG->TA point mutations (not shown), or poly dT are added. The only molecules that do not substantially quench are the natural RNA target site: the HIV RRE stem-loop IIB, and double-stranded poly dA-dT DNA, which is consistent with tryptophan stacking functioning as an important part of non-specific binding to DNA that is at least partly single-stranded, but not to dsDNA or structured RNA. However, under high salt conditions (Figure 6B), quenching by all the non-specific binders is significantly reduced or eliminated, while quenching by A3 and

29 is mostly unaffected. This strongly supports the idea that tryptophan stacking is a major component of specific, high affinity binding to these structures.

Finally, since the tryptophan is essential for structure recognition and is probably located at or very close to the junction of the DNA helices, we wanted to see if it was responsible for the peptide-induced conformational change described in Figure 4. To do this, we assayed the gel mobility of A3 and 29 when bound to the W45A mutant. As shown in Figure 7, we observed an ~30-fold increase in the concentration of peptide required to induce a gel shift, but the shift itself was in the same direction and of the same magnitude. This was also true when the R38A and R42A mutant peptides were assayed. Based on this, we conclude that the rearrangement of the helix orientation in response to Rev binding is not caused by a single peptide-DNA interaction at the helix junction, but instead, is determined by the whole peptide-DNA interface.

Discussion

Models for recognition of 3-helix junctions

The goal of our in vitro selection was to identify ligands that would serve as models for how ssDNA tertiary structure could contribute to binding affinity and specificity. The truncation data in Figure 2 indicates that the structure recognized by these DNA molecules is a 3-helix junction and that it plays a role in recognition that is quantitatively as important as the role of primary sequence. We have previously described a model for the recognition of the primary structure of these molecules (in preparation) and now we are interested in better understanding the mechanism of tertiary structure recognition.

There are several general strategies proteins might use to recognize branched nucleic acid helices[35]. They might simultaneously contact two of the helices in a specific orientation that is fixed by the branched structure. They might also recognize unpaired bases or unique features of the nucleic acid backbone which will certainly be present at the points of intersection between the helices. Finally, they might use aromatic sidechains to engage free helix ends, which are generally only accessible at helix branches, in stacking interactions. Examples of the first two modes of recognition have been characterized at atomic resolution (see below). Based on the biochemical experiments presented here, we propose that Rev uses stacking interactions as the primary mechanism of structure-specific binding.

The data in Figures 5 and 6 show that the tryptophan in Rev has the characteristics of a residue involved in base stacking- its aromaticity is essential and its fluorescence is dramatically quenched by DNA. Our previous work has shown that arginines 38 and 42

are likely to be in direct contact with the G•T/CG basestep of the TGTT helix. Since arginine 42 is ~1 turn of the α-helix away from the tryptophan, a distance of ~ 4.5 angstroms, and the G•T/CG basestep is one basepair, ~ 3.4 angstroms, from the end of the helix, the tryptophan should be located at the branchpoint, just beyond the end of the conserved helix. Given these surroundings, it is extremely likely that the tryptophan is stacked on the terminal basepair and/or the unpaired C residue. The ability of the phenylalanine and tyrosine substitutions to bind DNA suggest that it is exclusively the stacking function that produces the structure specificity and the inability of the alanine or even the leucine mutants to bind underscores the energetic importance of the interaction.

An element of this interaction which we have not directly addressed in this work is the relative orientation of the three helices. Studies with model 3- and 4-helix junctions suggest that the driving force of the folding of branched structures is coaxial stacking of helix pairs[36,37]. If we assume this is occurring in these molecules, it will be important to determine if the TGTT helix is part of a coaxial stack or whether it is the branched helix. If the TGTT helix is stacked, it would place the tryptophan between two helices, while it would abut only one helix if the TGTT was unstacked. It has been shown that the distance between basepairs in the B-form helix can be stretched enough to provide access to intercalators without disrupting basepairing, and the stacked basepairs at the junction in model 3-helix structures do exhibit a modestly increased helix rise [38]. However, stacking of the tryptophan with a branched TGTT helix rather than intercalation into a stacked helix would still probably require a smaller distortion of the junction and is thus more likely. Additionally, TGTT branching would explain the specificity of Rev for 3- rather than 2- or 4- helix junctions. Having an odd number of

helices assures that at least one will be unstacked and accessible to tryptophan, while an even number always leaves one helix to compete for the site occupied by tryptophan. In fact, our results suggest that 3-helix junctions generally may be excellent substrates for recognition by intercalation.

It is also possible that Rev does not require a specific helical arrangement or that the orientation varies in different molecules and depends on the sequence and structure of the helix linkers. Although this seems unlikely given the tight binding and strong selection for this fold, because of the wide variation in the size and sequence of the helix linkers, we cannot exclude the possibility. We have tried to use established electrophoresis techniques to determine the relationship of the three helices, but the large linkers between helices made the results difficult to interpret [25,26]. We are currently using NMR to more directly address this issue.

Whether the TGTT helix is stacked or not, the modification interference suggests that contacts are made to a second helix, which must be in close proximity. The peptide induced conformational change observed in the native gels suggests an acute angle between two helices which would be consistent with this. Mutagenesis has not indicated a specific role for these contacts and we have already assigned structural roles for all three of the essential amino acids. However, other arginine residues do show modest effects upon alanine substitution and it may be that loss of a single arginine that makes nonspecific contacts to the second helix can be compensated for by surrounding residues. Since this conformational change still occurs in the absence of the tryptophan, we believe that these non-specific interactions do contribute modestly to the affinity of the peptide for the DNA structure.

Comparison to other modes of branched helix recognition

In many ways, the mechanism used by Rev to recognize branched DNA is unique. As mentioned above, S15 is the best studied 3-helix binder and it recognizes the junction primarily by the backbone conformation on the outer surface of the structure. In contrast, where structures are available of Holliday junctions in complex with proteins involved in junction migration and resolution, most contacts are made along the lengths of the four helices by dimeric or tetrameric protein complexes, with junction specificity coming from a complementary arrangement between the individual DNA helices and protein binding sites[39-41]. Rev, however, uses the tryptophan to recognize the interior of the branchpoint. In some ways this is reminiscent of the proposed binding mechanism of the A-box from the HMG1 protein to 4-way DNA junctions [24,42]. In that situation, modeling has suggested that intercalation of a phenylalanine at the junction along with additional contacts in the minor groove of one arm of the junction accounts for the binding specificity. Intercalation, in fact, is common in many cases of sequence specific and non-specific DNA recognition by HMG and HMG-related proteins[43].

However, none of the HMG proteins that recognize 4-way junctions have the component of sequence specificity that Rev does and, interestingly, none are thought to use tryptophan for intercalation. This is surprising since tryptophan has been shown to function, based on the similarity of its shape to that of a purine base, as a "pseudobase" in the specific recognition of loop residues in the phage λ N peptide-box B RNA complex, as well as in stacking with ssDNA as part of non-specific ssDNA binding proteins[29,30,33]. This may be because the bulky sidechain can cause steric interference

at the core of a DNA:protein interface unless it is involved completely in a stacking interaction, a condition which could be difficult to satisfy unless the protein is, like Rev, deeply engaged in the DNA major groove.

This mode of junction recognition may be important because it provides a simple mechanism for recognizing the structure of 3-helix junctions, which occur in biologically specialized cases such as ssDNA viruses, and it might also apply to cruciforms, which are likely to be much more common in vivo. The Rev example suggests that the presence of an intercalating residue is enough to confer the ability to recognize helix junctions on motifs which recognize DNA sequence-specifically. This is beneficial for recognition of ssDNA in the cell because junctions are a necessary consequence of ssDNA extrusion from dsDNA, and they label even highly structured single-stranded sequences as ssDNA, providing a means to distinguish them from the vast excess of surrounding dsDNA. The fact that this strategy works in the context of an α-helix that binds in the major groove, which is the general mechanism of dsDNA recognition, also suggests that this mode of binding might easily be adopted by dsDNA binding proteins and that it might even be practiced by some proteins with aromatic residues near their DNA-binding surface.

Materials and Methods

Fluorescence-based binding assay

The procedures described in this section are adapted from the work of Luedtke and colleagues[44]. Rev peptide (Figure 1) with a C-terminal cysteine was fluorescein-labeled by incubating at 50μM with 500μM 5-(iodoacetamido)fluorescein (from a 10x stock in DMF) in 20mM sodium phosphate pH 8.0/2mM EDTA. The reaction was incubated at 25° for 2 hr. in the dark and purified by HPLC. Labeled peptide was detected and quantified by fluorescein absorbance at 475 nm.

For binding assays, fluorescein-labeled Rev (2.5nM) was mixed with DNA in 30mM Hepes pH 7.5/100mM KCl/ 40mM NaCl/ 10mM ammonium acetate/ 10mM guanidinium•HCl/ 2mM $MgCl_2$/ .5mM EDTA/ .001% Nonidet P-40. Binding reactions were 20μL in 384-well plates and were incubated for 30 min. Fluorescence intensities and anisotropies was measured in a LJL Biosystems Criterion fluorimeter with fluorescein filter sets (ex= 485nm, em=530 nm) and a G-factor of .8. All points were measured in quadruplicate and all values are the average of at least three independent experiments. Binding was measured by monitoring the DNA-dependent increase in peptide anisotropy. Binding data were fit to a single-site binding model using Kaleidagraph.

Competition binding assay

Competition assays were performed by incubating competitor peptide with 2.5 nM fluoresceinated Rev and 4nM A3 DNA. Reactions were allowed to equilibrate for 30

min. at 25° prior to measurement. Data were fit to a single site competition model and IC_{50} values were determined.

Gel shift assay

Gel shifts were performed in 10mM Hepes pH 7.5/100mM KCl/1mM $MgCl_2$/.5mM EDTA/50ug/mL yeast tRNA/10% glycerol at 4° using ^{32}P-end labeled DNA at a concentration of less than .5nM. Reactions were performed in 10μL volumes and incubated for 30 min., prior to separation on 10% polyacrylamide (37.5:1 mono:bis)/ .5xTBE gels. Gels were run 3 hr. at 225V.

Fluorescence quenching

Nucleic acids used for fluorescence measurements were isolated from 15% PAGE gels and electroeluted using the Elutrap system (Schleicher and Schuell) All measurements were made in degased at 25° in 10mM Tris pH 7.6/ /2mM $MgCl_2$,with either 20mM NacCl/50mM KCl or 80mM NaCl/200mM KCl, at peptide and nucleic acid concentrations of 2μM. Fluorescence spectra were taken on an ISS model K2 fluorimeter with a 450nM short pass excitation filter. The excitation wavelength was 295 nm and emission was recorded from 315-415nm. Experiments were repeated three times. Data was corrected for inner filter effects as described in [45].

References

1. Lilley, D. Structures of helical junctions in nucleic acids. *Quarterly Reviews of Biophysics* **33**, 109-159 (2000).

2. Lafontaine, D., Norman, D. & Lilley, D. The global structure of the VS ribozyme. *The EMBO Journal* **21**, 2461-2471 (2002).

3. Bassi, G., Murchie, A., Walter, F., Clegg, R. & Lilley, D. Ion-induced folding of the hammerhead ribozyme: a fluorescence resonance energy transfer study. *The EMBO Journal* **16**, 7481-7489 (1997).

4. Pearson, C. *et al.* Slipped-strand DNAs formed by long (CAG)•(CTG) repeats: slipped-out repeats and slip-out junctions. *Nucleic Acids Research* **30**, 4534-4547 (2002).

5. Pearson, C., Wang, Y.-H., Griffith, J. & Sinden, R. Structural analysis of slipped-strand DNA (S-DNA) formed in (CTG)n•(CAG)n repeats from the myotonic dystrophy locus. *Nucleic Acids Research* **26**, 816-823 (1998).

6. Pearson, C. & RR, S. Trinucleotide repeat DNA structures: dynamic mutations from dynamic DNA. *Current Opinion in Structural Biology* **8**, 321-330 (1998).

7. Ren, J. *et al.* Spectral and physical characterization of the inverted terminal repeat DNA structure from adenoassociated virus 2. *Nucleic Acids Research* **27**, 1985-1990 (1999).

8. Batey, R. & Williamson, J. Interaction of the Bacillus stearothermophilus ribosomal protein S15 with 16S rRNA: 1. Defining the minimal RNA site. *Journal of Molecular Biology* **261**, 536-549 (1996).

9. Batey, R. & Williamson, J. Interaction of the Bacillus Stearothermophilus ribosomal protein S15 with 16S rRNA: II. Specificity determinants of RNA-protein recognition. *Journal of Molecular Biology* **261**, 550-567 (1996).

10. Ha, T. *et al.* Ligand-induced conformational changes observed in single RNA molecules. *Proceedings of the National Academy of Sciences* **96**, 9077-9082 (1999).

11. Orr, J., Hagerman, P. & Williamson, J. Protein and Mg2+-induced conformational changes in the S15 binding site of 16S ribosomal RNA. *Journal of Molecular Biology* **275**, 453-464 (1998).

12. Agalarov, S., Prasad, G., Funke, P., Stout, C. & Williamson, J. Structure of the S15,S6,S18-rRNA complex: assembly of the 30S ribosome central domain. *Science* **288**, 107-112 (2000).

13. Spiro, C., Bazett-Jones, D., Wu, X. & McMurray, C. DNA structure determines protein binding and transcriptional efficiency of the proenkephalin cAMP-responsive enhancer. *Journal of Biological Chemistry* **270**, 27702-27710 (1995).

14. Spiro, C. & McMurray, C. Switching of DNA secondary structure in proenkephalin transcriptional regulation. *Journal of Biological Chemistry* **272**, 33145-33152 (1997).

15. Glucksmann-Kuis, M.A., Dai, X., Markiewicz, P. & Rothman-Denes, L.B. E. coli SSB activates N4 virion RNA polymerase promoters by stabilizing a DNA hairpin required for promoter recognition. *Cell* **84**, 147-154 (1996).

16. Liu, B., Maul, R. & Kaetzel, D. Repression of platelet-derived growth factor A-chain gene transcription by an upstream silencer element. *Journal of Biological Chemistry* **271**, 26281-26290 (1996).

17. Thakur, S. *et al.* Regulation of BRCA1 transcription by specific single-stranded DNA binding factors. *Molecular and Cellular Biology* **23**, 3774-3787 (2003).

18. Desveaux, D., Despres, C., Joyeaux, A., Subramaniam, R. & Brisson, N. PBF-2 is a novel single-stranded DNA binding factor implicated in PR-10a gene activation in potato. *The Plant Cell* **12**, 1477-1489 (2000).

19. Becker, N., Kelm, R., Vrana, J., Getz, M. & Maher, L. Altered sensitivity to single-strand specific reagents associated with the genomic vascular smooth muscle α-actin promoter during myofibroblast differentiation. *Journal of Biological Chemistry* **275**, 15384-15391 (2000).

20. Basar, L. *et al.* Targeted melting and binding of a DNA regulatory element by a transactivator of c-myc. *Journal of Biological Chemistry* **270**, 8241-8248 (1995).

21. Michelotti, G. *et al.* Multiple single-stranded cis elements are associated with activated chromatin of the human c-myc gene in vivo. *Molecular and Cellular Biology* **16**, 2656-2669 (1996).

22. Farokhzad, O., Teodoridis, J., Park, H., Arnaout, M. & Shelley, C. CD43 gene expression is mediated by a nuclear factor which binds pyrimidine-rich single-stranded DNA. *Nucleic Acids Research* **28**, 2256-2267 (2000).

23. Zuker, M. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Research* **31**, 3406-3415 (2003).

24. Webb, M. & Thomas, J. Structure-specific binding of the two tandem HMG boxes of HMG1 to four-way junction DNA is mediated by the A domain. *Journal of Molecular Biology* **294**, 373-387 (1999).

25. Welch, J., Walter, F. & Lilley, D. Two inequivalent folding isomers of the three-way DNA junctions with unpaired bases: sequence-dependence of the folded conformation. *Journal of Molecular Biology* **251**, 507-519 (1995).

26. Welch, J., Duckett, D. & Lilley, D. Structures of bulged three-way DNA junctions. *Nucleic Acids Research* **21**, 4548-4555 (1993).

27. Escalante, C., Yie, J., Thanos, D. & Aggarwal, A. Structure of IRF-1 with bound DNA reveals determinants of interferon regulation. *Nature* **391**, 103-106 (1998).

28. Shakked, Z. *et al.* Determinants of repressor/operator recognition from the structure of the trp operator binding site. *Nature* **368**, 469-473 (1994).

29. Legault, P., Li, J., Mogridge, J., Kay, L.E. & Greenblatt, J. NMR structure of the bacteriophage lambda N peptide/boxB RNA complex: recognition of a GNRA fold by an arginine-rich motif. *Cell* **93**, 289-299 (1998).

30. Tan, R., Chen, L., Buettner, J.A., Hudson, D. & Frankel, A.D. RNA recognition by an isolated α helix. *Cell* **73**, 1031-1040 (1993).

31. Lohman, T. & Overman, L. Two binding modes in Escherichia coli single strand binding protein-single-stranded DNA complexes. *Journal of Biological Chemistry* **260**, 3594-3603 (1985).

32. Mely, Y. *et al.* Binding of the HIV-1 nucleocapsid protein to the primer $tRNA^{Lys}$ in vitro, is essentially not specific. *Journal of Biological Chemistry* **270**, 1650-1656 (1995).

33. Raghunathan, S., Kozlov, A., Lohman, T. & Waksman, G. Structure of the DNA binding domain of E.coli SSB bound to ssDNA. *Nature Structural Biology* **7**, 648-652 (2000).

34. Behmoaras, T., Toulme, J. & Helene, C. Specific recognition of apurinic sites in DNA by a tryptophan-containing peptide. *Proceedings of the National Academy of Sciences of the United States of America* **78**, 926-930 (1981).

35. Duckett, D., Murchie, A., Giraud-Panis, M., Pohler, J. & Lilley, D. Structure of the 4-way junction and its interaction with proteins. *Philosophical Transactions of the Royal Society of London* **347**, 27-36 (1995).

36. Ouporov, I. & Leontis, N. Refinement of the solution structure of a branched DNA three-way junction. *Biophysical Journal* **68**, 266-274 (1995).

37. van Buuren, B., Overmars, F., Ippel, J., Altona, C. & Wijmenga, S. Solution structure of a DNA three-way junction containing two unpaired thymidine bases. Identification of sequence features that decide conformer selection. *Journal of Molecular Biology* **304**, 371-383 (2000).

38. Yang, X.-l., Robinson, H., Gao, Y.-G. & Wang, A.-J. Binding of a macrocyclic bisacridine and ametantrone to CGTACG involves similar unusual intercalation platforms. *Biochemistry* **39**, 10950-10957 (2000).

39. Ariyoshi, M., Nishino, T., Iwasaki, H., Shinagawa, H. & Morikawa, K. Crystal structure of the Holliday junction DNA in complex with a single RuvA tetramer. *Proceedings of the National Academy of Sciences* **97**, 8257-8262 (2000).

40. Roe, S. *et al.* Crystal structure of an octameric RuvA-Holliday junction complex. *Molecular Cell* **2**, 361-372 (1998).

41. Declais, A.-C. *et al.* The complex between a four-way DNA junction and T7 endonuclease I. *The EMBO Journal* **22**, 1398-1409 (2003).

42. Thomas, J. & Travers, A. HMG1 and 2, and related architectural DNA-binding proteins. *Trends in Biochemical Sciences* **26**, 167-174 (2001).

43. Bewley, C., Gronenborn, A. & Clore, G. Minor groove binding architectural proteins: structure, function, and DNA recognition. *Annual Review of Biophysics and Biomolecular Structure* **27**, 105-131 (1998).

44. Luedtke, N., Liu, Q. & Tor, Y. RNA-ligand interactions: affinity and specificity of aminoglycoside dimers and acridine conjugates to the HIV-1 Rev response element. *Biochemistry* **42**, 11391-11403 (2003).

45. Lohman, T. & Mascotti, D. Thermodynamics of ligand-nucleic acid interactions. *Methods in Enzymology* **212**(1992).

Figure legends

Figure 1. Proposed ssDNA molecules that bind Rev. (A) Predicted secondary structures of sequences A3 (left) and 29 (right). Positions where chemical modification strongly reduces binding are in bold red capitals while those that moderately reduce binding are in bold black capitals. Box indicates conserved G•T/CG basestep that is essential for sequence-specific Rev binding. (B) Sequence of the Rev peptide used in this paper. Residues in large type are positions which show at least a 10-fold loss in affinity when mutated to alanine.

Figure 2. Helix truncations eliminate high affinity Rev binding. (A) Representative binding curves, showing binding of sequence 29 (—●—), 29AB1 (—◆—), 29AB2 (— ■ -), 29BC1 (—▲—), and 29BC2 (— ▼ -). (B) Diagrams of truncation mutants and binding affinities relative to parent 3-helix molecule.

Figure 3. Cruciforms bind Rev with high affinity but relaxed specificity. (A) Binding curves for molecules A3 (—▲—), A3 CG->AT (— ▼ -), A3 cruciform 1 (—■—), A3 cruciform 1 CG->AT (— ◆ -), A3 cruciform 2 (—●—) and A3 cruciform 2 CG->AT (— ⊞ -). (B) Summary of binding data. Affinities are relative to the A3 parent and specificity is the ratio of each molecule to that of the same molecule with a CG->AT mutation, as indicated by asterisks.

Figure 4. Rev causes an increase in the gel mobility of 3-helix molecules. Gel shifts for 29 and A3. Numbers are concentrations of Rev peptide.

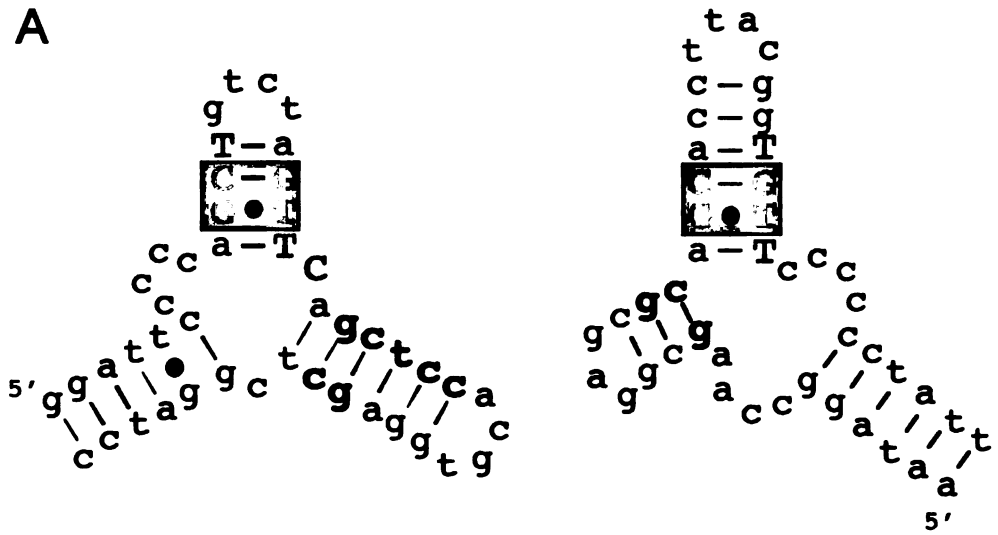
Figure 5. Recognition depends on an aromatic residue at position 45. Competition binding curves with fluorescein labeled Rev and indicated amounts of unlabeled mutant competitor. Competitors are unlabeled Rev (—●—), Rev W45A (—■—), Rev W45F(—▲—), Rev W45L (—▼—), and Rev W45Y (—▼—).

Figure 6. Tryptophan fluorescence is quenched in specific complexes. Fluorescence spectra ($\lambda_{excitation}$=295) (a) under low salt conditions (50mM KCl/20mM NaCl) and (b) high salt conditions (200mM KCl/80mM NaCl). Molecules are Rev alone (—⊞—), + A3 (—●—), +A3BC1 (—○—), +29 (—▲—), +29BC1 (—△—), +IIB (—×—), +15-mer dT (—◆—), and +15-mer dAdT (—◇—). Truncation mutants are from Figure 2.

Figure 7. Rev induces a conformational change in the DNA even without tryptophan. Gel shifts are done with A3 DNA and Rev (top), Rev R42A (middle), and Rev W45A (bottom). Numbers indicate concentration of peptide in each lane.

Figure 1

A



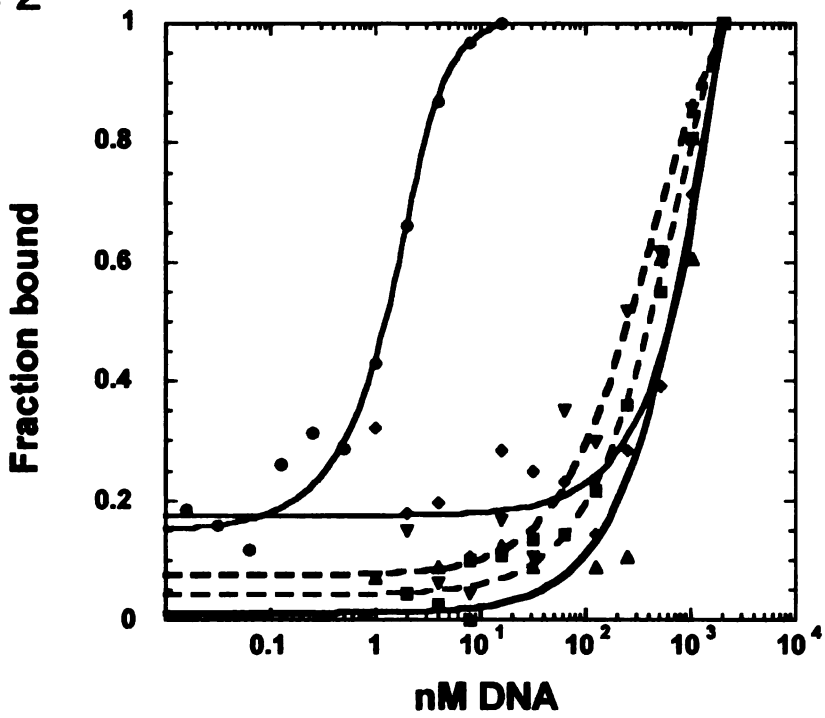
A3 29

B

$_{34}$SUCCTRQA**R**RNR**R**RR**W**RERQR$_{50}$AAAAR

Figure 2

A


Fraction bound
1
0.8
0.6
0.4
0.2
0
0.1
1
10 1
10 2
10 3
10 4
nM DNA

B

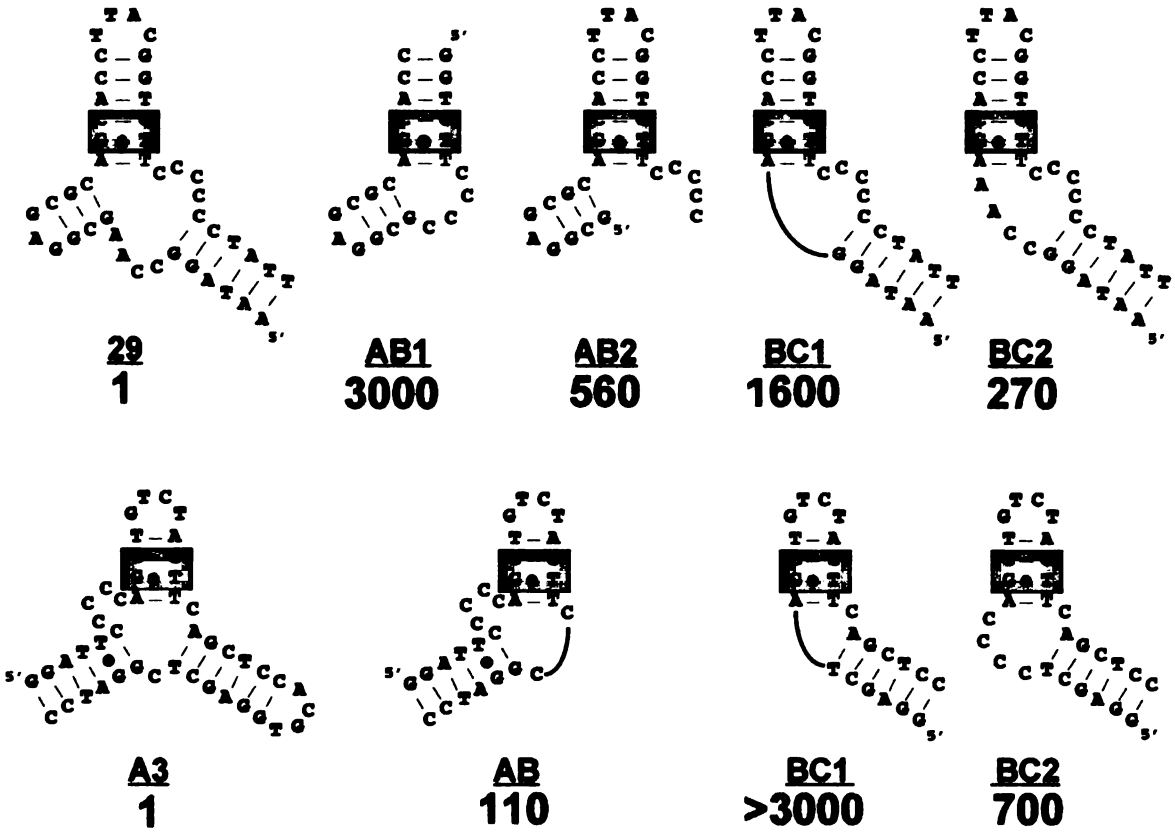
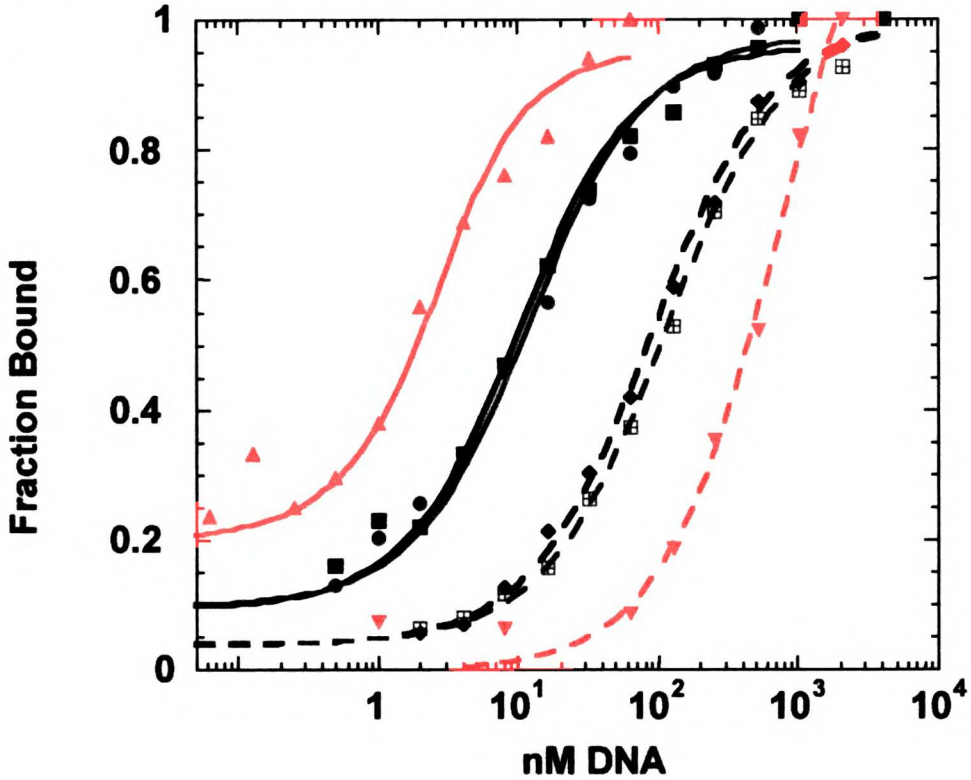
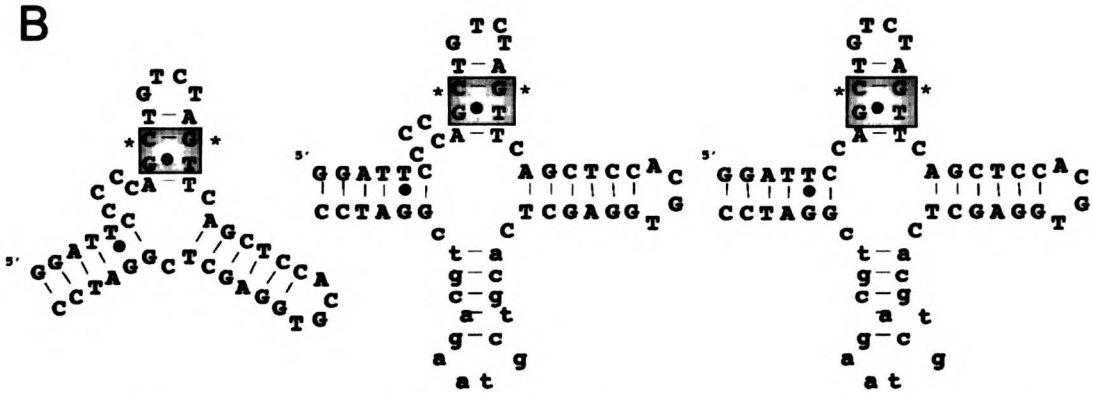

29
1
AB1
3000
AB2
560
BC1
1600
BC2
270
A3
1
AB
110
BC1
>3000
BC2
700

Figure 3

A



B



	A3	A3 Cruciform 1	A3 Cruciform 2
K_{rel}:	1	4.0	3.9
Specificity:	335	8.6	10.3

Figure 4

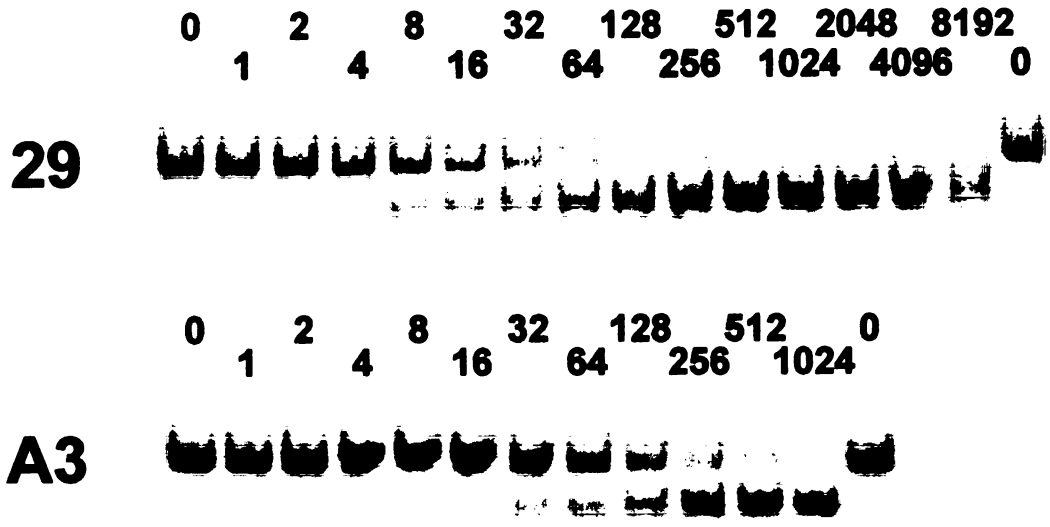

0 2 8 32 128 512 2048 8192
1 4 16 64 256 1024 4096 0
29
0 2 8 32 128 512 0
1 4 16 64 256 1024
A3

Figure 5

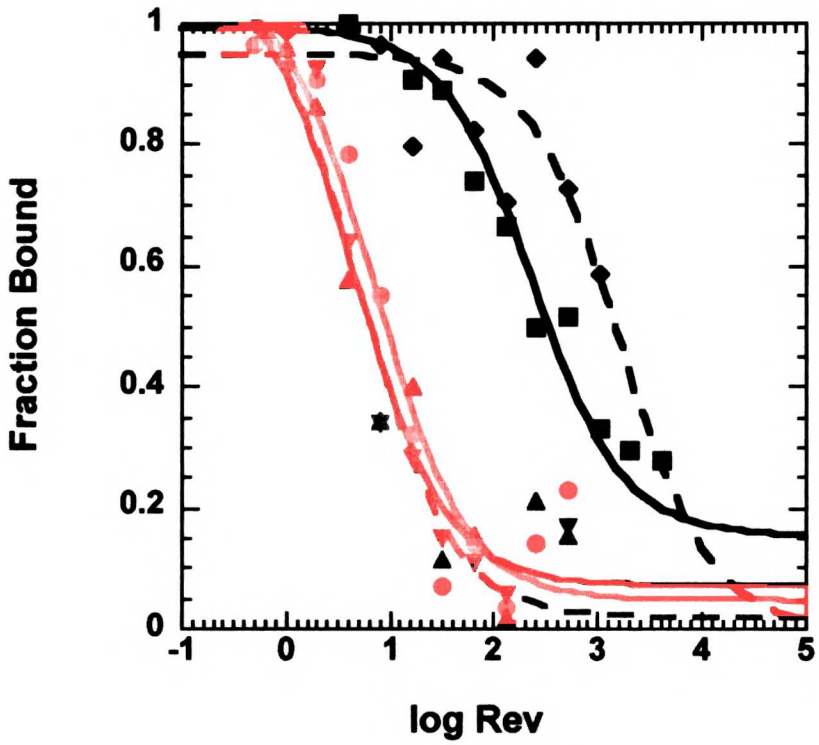


Figure 6

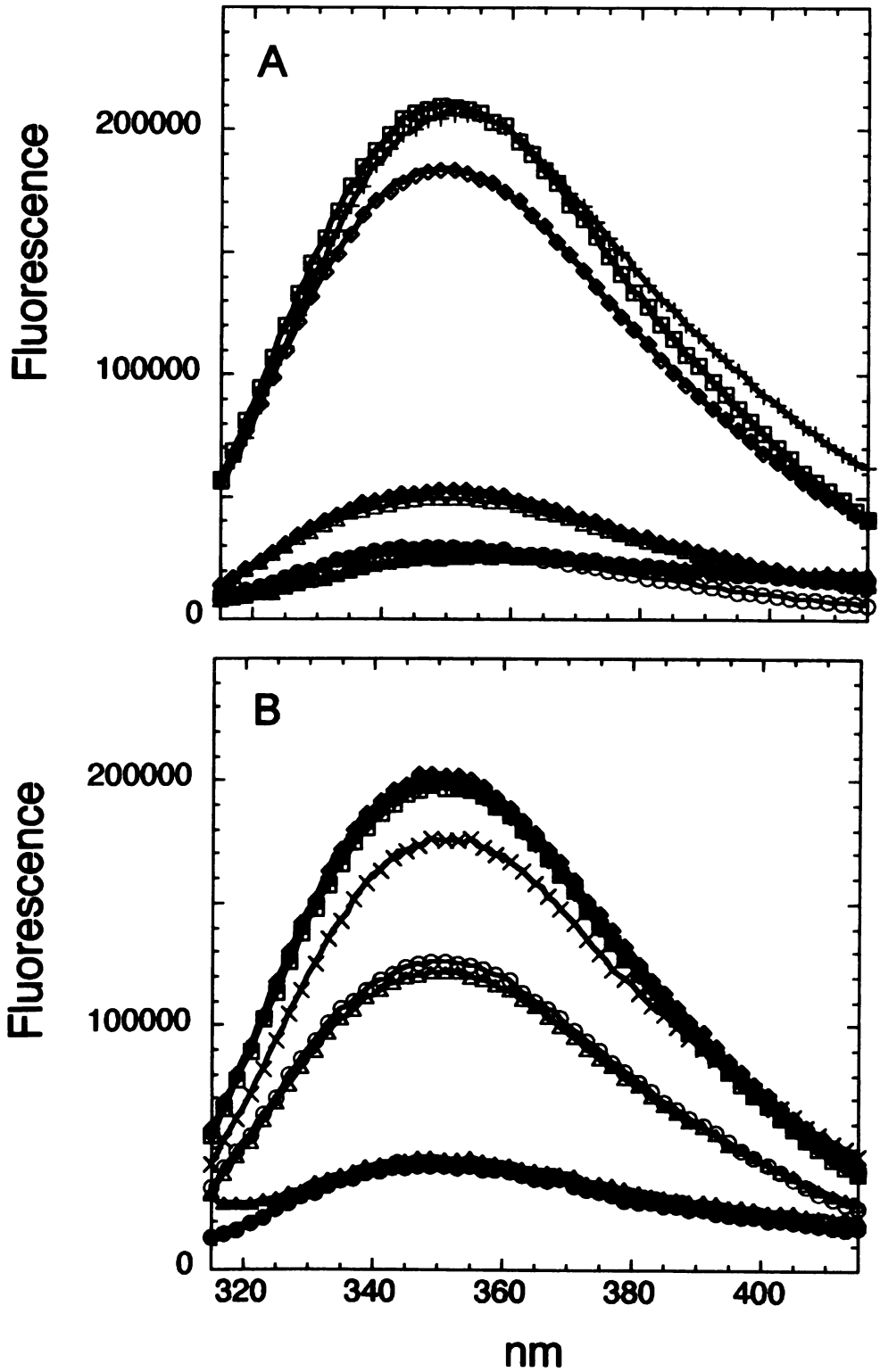
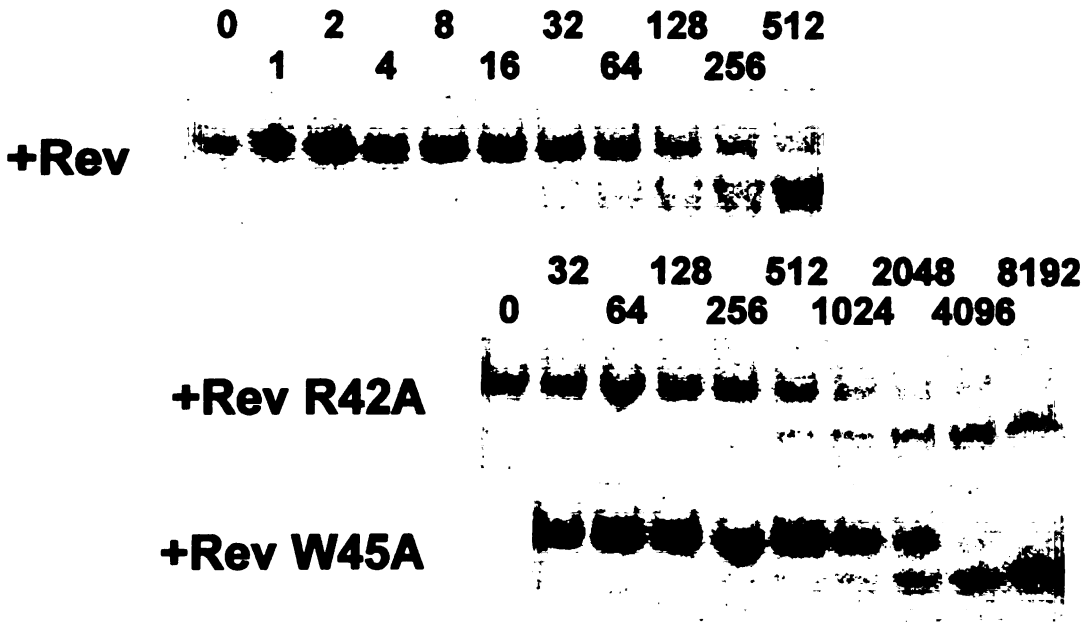


Figure 7


0 2 8 32 128 512
1 4 16 64 256
+Rev
32 128 512 2048 8192
0 64 256 1024 4096
+Rev R42A
+Rev W45A

Appendix 1

Determination of minimal binding regions in selected ssDNA molecules

One of the goals of the in vitro selection described in Chapter 3 was to assess the role of ssDNA tertiary structure in promoting protein recognition. In choosing Rev as the protein ligand, we expected to select for heavily structured molecules based on the analogy to its RNA binding site and the RNA sites of other arginine-rich motif peptides. However, because the B-form DNA helix geometry seems too wide and shallow to allow a large enough interface for tight binding by a single α-helix, we suspected that molecules with relatively complex tertiary structures, beyond simple bulges or loops, would be favored. After computationally folding the full-length sequenced molecules, we were excited to see such an outcome, reflected in the putative 3-helix structure in the most abundant motif selected. The R28 molecule was also interesting because it displayed some similarity to another orphan sequence in a region that was predicted to form an almost perfect hairpin. However, this speculation was based entirely on the computational folds and there were structurally distinct but energetically similar folds for all molecules.

To better understand the secondary structures, one of the first characterization experiments I performed was a boundary analysis. This involves using limited nuclease digestion to fragment the full length DNA molecules into a pool of different sizes. The digested DNA is then bound to the peptide, isolated, and material that binds is compared to the unfractionated input. Molecules can be labeled at either the 5' or 3' end to identify the 3' and 5' boundaries respectively. Results are shown in Figure 1 with a summary superimposed on the suggested folds in Figure 2.

For all the molecules that I tested that contain the conserved TGTTC/AGCA helix, the observed boundaries are consistent with an essential role for the predicted 3-helix

junction (Figure 2a). For example, in figure 1a, it is clear that little if any deletion from the 3' end of the A3 molecule is tolerated, while figure 1b shows that approximately half the sequence from the 5' end can be deleted. The only potentially ambiguous case is molecule 29, which appears to not require elements of the third helix. However, the 3' boundary falls in a run of six consecutive C residues and the boundary suggests a requirement for approximately three of them. It is likely, therefore, that when the C residues that basepair to complete the 3-helix junction in the fold shown in Figure 2a are removed, the others can be used to close the structure. This would suggest that the apparently large poly-C linker shown in Figure 2a is mostly dispensable.

For molecule R28, the observed boundaries are entirely consistent with a short, single hairpin binding site, as described extensively in Chapter 3.

Materials and Methods

DNA molecules were 5' end-labeled with T4 polynucleotide kinase (New England BioLabs) and γ-^{32}P-ATP according to manufacturer's instructions. 3'-end labelings were done using 20 units of terminal transferase (New England BioLabs) and 4 pmol α-^{32}P cordycepin for 4 pmol DNA. Both labelings were twice purified using a Qiagen nucleotide removal kit.

4 pmol of labeled DNA was digested using 2 units of S1 nuclease (Promega) for 2 min. at 25° in the presence of 2.5μg salmon sperm carrier DNA. Reactions were ethanol precipitated twice and resuspended in binding buffer. Binding buffer and DNA fractionation are exactly as described in the modification interference section of the

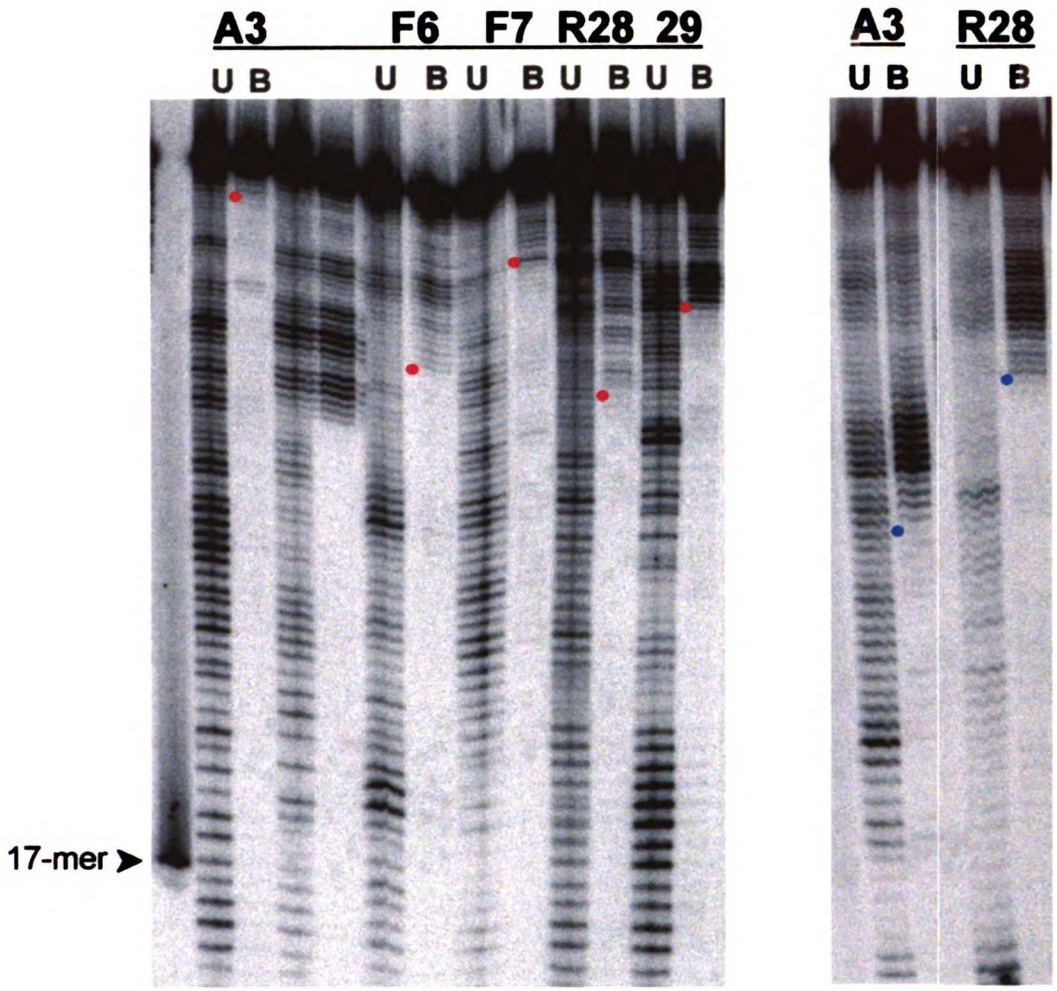


Figure 1. Binding site boundary determination of selected Rev-binding DNA molecules. Representative gels showing unfractionated DNA (U) and Rev bound DNA (B) for the indicated sequences. Dots indicate boundaries (see Figure 2). DNA is 5'-end labeled in left panel, to allow determination of the 3' boundary and 3'-end labeled in right panel, to allow determination of the 5' boundary. A 17-mer sequence used as a size marker is indicated.

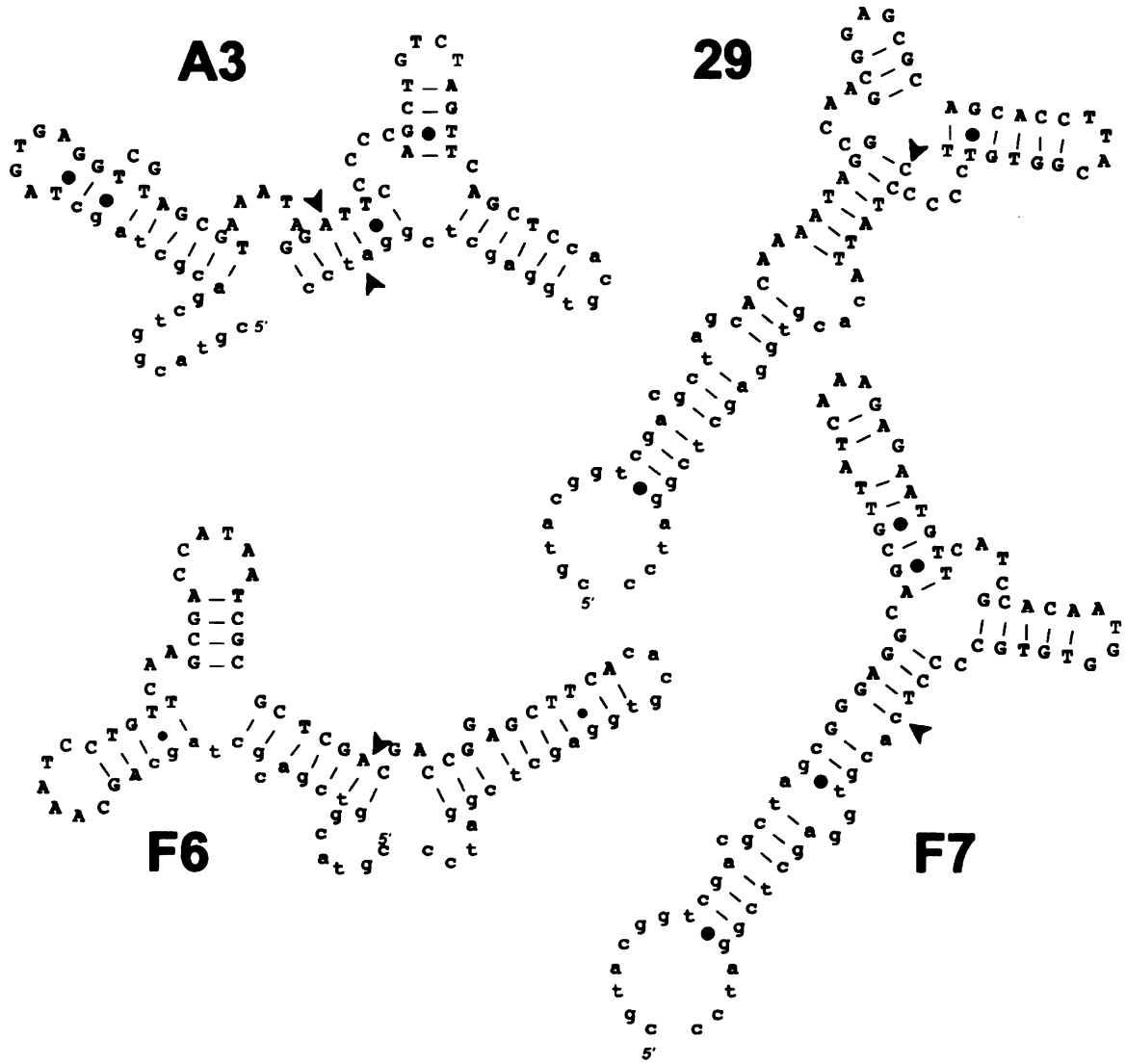


Figure 2. Rev-binding site boundaries in selected DNA molecules of the TGTTC/AGCA class. Superimposition of boundaries predicted from Figure 1 on the most plausible predicted secondary structures. Red arrows indicate 3' boundaries and blue arrows indicate 5' boundaries from Figure 1.

R28

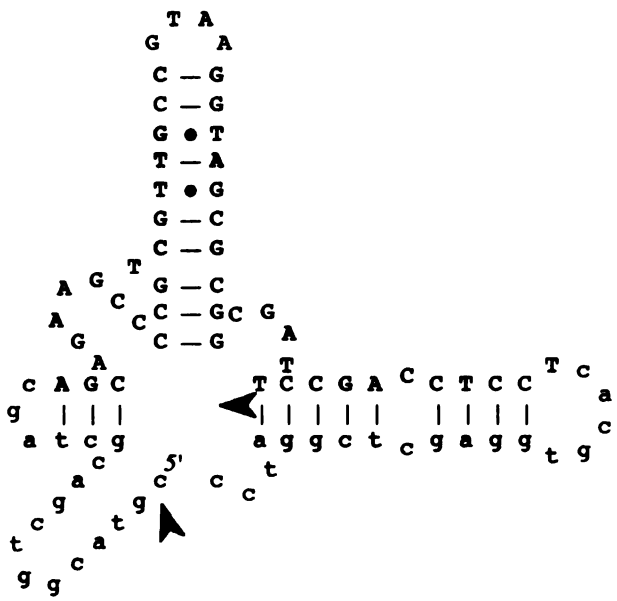


Figure 3. Rev-binding site boundaries in the R28 molecule. Superimposition of boundaries predicted from Figure 1 on the most plausible predicted secondary structure of R28. Arrows are as for Figure 2.

Appendix 2

Salt dependence of Rev:DNA and Rev: RNA interactions

A source of affinity, but not necessarily specificity, for any protein:nucleic acid complex is electrostatic interaction between basic amino acids and the negatively charged phosphate backbone of DNA or RNA. Obviously this will be the case with Rev (see Fig. 1a, Chapters 3 and Chapter 4) or any other arginine-rich peptide:nucleic acid interaction. It is interesting, however, that Rev appears, based on the alanine scanning mutagenesis results, to use fewer amino acids to bind the selected DNA molecules with high affinity than it does to bind RNA. Four residues were identified as essential for R28 recognition, and only three are necessary for A3/3-helix recognition, but no fewer than six residues are important for interacting with RRE IIB RNA (see Chapter 3, Figure 5)[1]. One explanation for this difference is that the DNA:peptide complexes may rely more on electrostatic interactions. In a peptide containing 11 arginines, this might not be apparent in the context of single arginine->alanine substitutions, where neighboring arginines might compensate for the loss of a single positive charge.

To test this, we assayed the dependence of Rev binding to R28 DNA, to A3/3-helix DNA, and to RRE IIB RNA on the concentration of ions in solution. In polyelectrolyte theory, the slope of a log/log plot of the binding constant against the salt concentration is taken to reflect the number of ions released in a binding event. The results are shown in the accompanying figure.

It is apparent that binding of both IIB and R28 show similar sensitivities to salt concentration, while binding of the 3-helix molecules is significantly less sensitive. For IIB and R28, the slopes are ~6.4, reflecting the release of ~6 ions upon complexation, while the slope is 4.5 for the A3 molecule, indicating the release of ~4 ions [2]. The value obtained for IIB is consistent with the observation from the NMR structure that six

arginines are in close enough proximity to the RNA to directly make electrostatic contacts with the backbone [3,4]. This is also consistent with the mutagenesis, which identified 4 essential arginines, in addition to 3 others which function when mutated to lysine, but have show a moderate defect when mutated to alanine[5].

These results do not explain the limited number of essential amino acids needed for tight DNA binding. Rather, they suggest, for A3, that electrostatics are less important than they are for RNA recognition. This may partly be expected because one of the stringency controls in our experiment was to eliminate molecules with an affinity for a Rev peptide in which all the arginines have been mutated to lysine, selecting against molecules which relied heavily on electrostatics. However, it does further underscore the quality of the individual interactions that have been selected for: the di-arginine-G•T/CG interaction (Chapter 3) and the tryptophan stacking interaction (Chapter 4). It also demonstrates that, while ARM peptides will certainly use electrostatics as a means of binding nucleic acid, it is far from the most important strategy for recognition.

Finally, it should be pointed out that the slope of the lines in the plot can be used to extrapolate binding constants for these molecules under more physiologically reasonable salt conditions. At 100mM NaCl, the binding constant for A3 is predicted to be ~ 70 pM, while for R28, it should be ~3 pM. These are conditions similar to those under which DNA binding proteins are often measured and these affinities are higher than those generally measured for DNA binding[6-8]. This further attests to the value of ssDNA recognition

Materials and Methods

Binding assays were carried out exactly as described in Materials and Methods, Chapter3. Binding buffer was 10 mM Hepes pH 7.5/2mM $MgCl_2$/.5mM EDTA/.001% Nonidet P-40, with NaCl at a range of concentrations between 120mM and 600mM.

References

1. Tan, R., Chen, L., Buettner, J.A., Hudson, D. & Frankel, A.D. RNA recognition by an isolated α helix. *Cell* **73**, 1031-1040 (1993).

2. Lohman, T. & Mascotti, D. Thermodynamics of ligand-nucleic acid interactions. *Methods in Enzymology* **212**(1992).

3. Grate, D. & Wilson, C. Role REVersal: understanding how RRE RNA binds its peptide ligand. *Structure* **5**, 7-11 (1995).

4. Battiste, J.L. *et al.* Alpha helix major groove recognition in an HIV-1 Rev peptide-RRE RNA complex. *Science* **273**, 1547-1551 (1996).

5. Tan, R. & Frankel, A. Costabilization of peptide and RNA structure in an HIV Rev peptide-RRE complex. *Biochemistry* **33**, 14579-14585 (1994).

6. Chin, J. & Schepartz, A. Concerted evolution of structure and function in a miniature protein. *Journal of the American Chemical Society* **123**, 2929-2930 (2001).

7. Montclare, J. & Schepartz, A. Miniature homeodomains: high specificity without an N-terminal arm. *Journal of the American Chemical Society* **125**, 3416-3417 (2003).

8. Zondlo, N. & Schepartz, A. Highly specific DNA recognition by a designed miniature protein. *Journal of the American Chemical Society* **121**, 6938-6939 (1999).

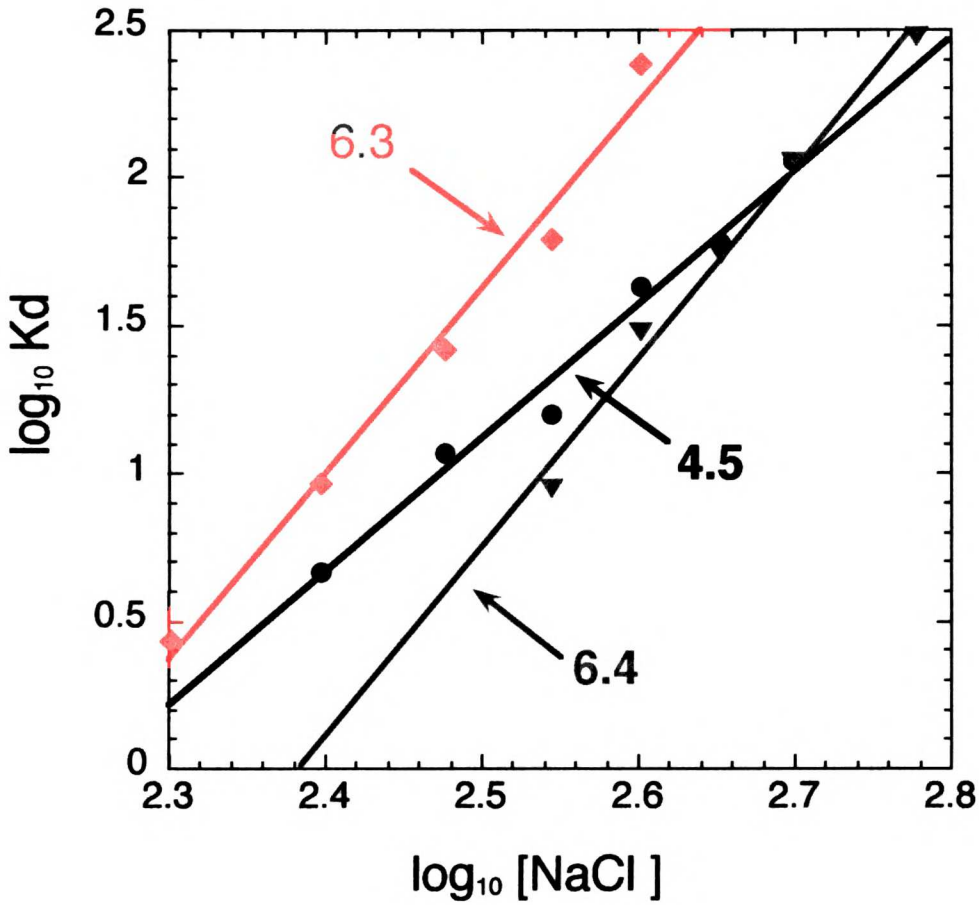


Figure 1: Salt dependence of Rev binding to DNA and RNA ligands. Log-log plot of salt dependence of Rev binding to A3 DNA (—●—), R28 DNA (—▼—), and RRE IIB RNA (—◆—). Slopes of best linear fits are indicated on the plot. Kd values are in nM units, [NaCl] values are in mM units.

Appendix 3

Analysis of helix conformation in selected 3-helix junctions

The motivation of much of the work presented in this thesis has been to understand the roles tertiary structure can play in the recognition of ssDNA. Because of this, the objective that emerged from the discovery that Rev recognizes a 3-helix junction, and the focus of Chapter 4, is to understand what facet of this structure is recognized. One central observation that has been made based on the study of naturally-occurring and model 3-helix junctions is that coaxial stacking of two of the helices is a driving force in the folding of these structures [1-6]. That this sort of effect has a role in Rev recognition is suggested by the observation in Figure 4, Chapter 4 that peptide binding causes an increase in the gel mobility of 3-helix junction sequences A3 and 29. The most reasonable interpretation of this is that Rev causes the DNA to move to a conformation in which two of the helices are stacked. This creates a long axis in the molecule that can align itself with the electric field, allowing the DNA to move through the gel matrix with only the single unstacked helix projecting outward to impede its progress, rather than having two helices stick out when the length of the third is aligned with the electric field. In the absence of Rev, the DNA may either be unstacked and in a Y-shaped conformation or perhaps in a conformation with an alternate stacking pattern.

A potentially important question that emerges is: what are the stacking patterns in the 3-helix junctions that bind Rev? To understand if this plays a role in structure-specific recognition, we also need to know if this pattern is a conserved feature of the selected molecules. Additionally, if we know whether the directly contacted TGTT/AGCA helix is in a stacked or unstacked conformation, we can deduce whether the essential tryptophan in Rev is likely to be stacking between the ends of one or two helices.

A technique to determine the orientation of helices in a branched structure has been developed by Lilley and colleagues and used to study model 3-helix junctions containing formally unpaired nucleotides between only one pair of the helices [3,4]. This method involves extending the lengths of all three helices to ~40 nucleotides each and incorporating unique restriction sites in each helix near the junction. Each helix can then be independently shortened. The relative gel mobilities of the truncated structures have been interpreted to be proportional to the angle subtended by the two undigested helix arms. In this assay, shortening of the unstacked helix is expected to result in the greatest increase in mobility, while digestion of either stacked arm should have a smaller effect.

I have attempted this analysis on two of the selected molecules, sequences A3 and 29. In each case, I have introduced HindIII, EcoRI, and XbaI sites into the three arms, respectively, with the conserved TGTT/AGCA helix containing the Eco site, while the Hind site is in the arm 5' of the TGTT helix and the Xba arm is 3' of the TGTT helix. The results of one gel, run in the absence of mono- or divalent cations, is shown in the accompanying figure, along with a schematic interpretation of the branching patterns.

For A3, the pattern seen in the absence of Rev, where only Xba digestion dramatically increases DNA mobility, is consistent with the helix 3' to the TGTT helix branching from a stack involving the other two helices. When Rev is added, a pattern appears in which both Hind and Xba digestion increase mobility. This can be interpreted in two ways. It is possible that a conformation is present in which the TGTT helix forms an obtuse angle with both of the other helices, which are related to each other by an acute angle. Alternately, there may be two conformations in rapid equilibrium in which the TGTT helix stacks with either the 5' or the 3' helix, resulting in a mobility that is an average of

the mobilities of the two species. While this gel does not make it possible to distinguish between these two possibilities, it should be noted that the HindIII and XbaI digests in the presence of 250nM Rev demonstrates that distinct species, representing the bound and unbound states, can be resolved. This suggests that different species are long-lived relative to the time scale of the gel and argues against the rapid equilibrium model

For 29, the migration pattern in the absence of Rev suggests a conformation in which the TGTT helix is stacked with its 3' neighbor, which is related by an obtuse angle to the helix 5' of the TGTT. In the presence of Rev, there is a downward shift in the mobility of the Eco digestion product, suggesting a movement of the TGTT helix to a slightly less stacked conformation and coaxial stacking between the remaining helices.

A tentative conclusion of this analysis is that, in the bound state, the conserved TGTT helix appears to be imperfectly stacked and is less stacked than in the free state. This would suggest that the tryptophan in Rev stacks on the free end of this helix alone. However, there are several potential problems with this interpretation. First, the peptide induced gel shifts only occur at very high Rev concentrations, ~ 250 nM for A3 and ~1.5 μM for 29. These concentrations are between 5- and 50-fold higher than those required to cause an increase in the mobility of the DNAs in their natural sequence context (Figure 4, Chapter 4), which raises questions about the meaning of the shift seen here. Second, these predicted stacking patterns are unexpected in the context of the predicted fold for sequence 29. In this case, the only two helices with no predicted unpaired bases between them are suggested by this data to be related by the smallest angle, which is inconsistent with the results of studies model 3-helix junctions [3,4]. Finally, in the case of sequence 29, if the conformational change predicted here is in fact

occurring, it seems surprising that there would be a downward mobility shift in the natural context since a predicted short unstacked helix in the free state is being exchanged for a predicted longer one in the bound state. While the results and conclusions described here may be valid, it is also possible that problems may have arisen because of inaccurate predictions of the folds near the junctions or because the rules for interpreting the mobility shifts derived from model junctions do not hold for molecules such as these, with multiple unpaired bases between several of the helices. Additional experiments, either using NMR or fluorescence resonance energy transfer between fluorescently tagged ends of helices will be needed to satisfactorily determine the relationships between the helical arms of these molecules.

Materials and Methods

For each 3-helix junction, three oligonucleotides were designed with the core junction sequences from selected molecules A3 or 29 surrounded by the helical arm sequences used previously to characterize model junctions[4]. For A3, these sequences were:

5'-CGCAAGCGACAGGAACCTCGAGAAGCTTCCGGTAGC**ATTCCCCCAGCT**
CGGTGGTTGAATTCCTCGAGGTTCCTGTCGCTTGCG

5'-CGCAAGCGACAGGAACCTCGAGGAATTCAACCACCG**AGTTCAGCT**
AACTGCAGTCTAGACTCGAGGTTCCTGTCGCTTGCG

5'-CGCAAGCGACAGGAACCTCGAGTCTAGACTGCAGTT**AGCTCGGAT**
GCTACCGGAAGCTTCTCGAGGTTCCTGTCGCTTGCG

For 29, these sequences were

5'-CGCAAGCGACAGGAACCTCGAGAAGCTTCCGGTAGC**CGCAGCA**

CGGTGGTTGAATTCCTCGAGGTTCCTGTCGCTTGCG

5'-CGCAAGCGACAGGAACCTCGAGGAATTCAACCACCG**TGTTCCCCCCTA**

AACTGCAGTCTAGACTCGAGGTTCCTGTCGCTTGCG

5'-CGCAAGCGACAGGAACCTCGAGTCTAGACTGCAGTT**TAGGCCAAGCG**

GCTACCGGAAGCTTCTCGAGGTTCCTGTCGCTTGCG

(bold indicates junction sequence from A3 or 29, underlines indicate restriction sites).

Molecules were end-labeled and Qiagen purified as described in Materials and Methods, Chapter 3. Molecules were annealed in 80μL at 50nM in 1x EcoRI buffer (New England BioLabs) in the presence of trace amounts of one of the labeled strands. Annealing reactions were digested with 60 units of the appropriate enzyme for 30 min. at 37°, followed by heat inactivation for 20 min. at 70° and slow cooling to room temperature. 3.5μL of each reaction was diluted with 3μL 4X gel shift buffer (40mM Hepes pH 7.5/400mM KCl/4mM $MgCl_2$/2mM EDTA/200μg/mL yeast tRNA/40% glycerol) to a total volume of 12μL and allowed to equilibrate for 30 min. at 25° in the presence or absence of Rev. Bands were resolved on 8% acrylamide (19:1) in 1xTBE run overnight at 375V.

References

1. Orr, J., Hagerman, P. & Williamson, J. Protein and Mg2+-induced conformational changes in the S15 binding site of 16S ribosomal RNA. *Journal of Molecular Biology* **275**, 453-464 (1998).

2. Batey, R. & Williamson, J. Interaction of the Bacillus stearothermophilus ribosomal protein S15 with 16S rRNA: 1. Defining the minimal RNA site. *Journal of Molecular Biology* **261**, 536-549 (1996).

3. Welch, J., Duckett, D. & Lilley, D. Structures of bulged three-way DNA junctions. *Nucleic Acids Research* **21**, 4548-4555 (1993).

4. Welch, J., Walter, F. & Lilley, D. Two inequivalent folding isomers of the three-way DNA junctions with unpaired bases: sequence-dependence of the folded conformation. *Journal of Molecular Biology* **251**, 507-519 (1995).

5. van Buuren, B., Overmars, F., Ippel, J., Altona, C. & Wijmenga, S. Solution structure of a DNA three-way junction containing two unpaired thymidine bases. Identification of sequence features that decide conformer selection. *Journal of Molecular Biology* **304**, 371-383 (2000).

6. Ouporov, I. & Leontis, N. Refinement of the solution structure of a branched DNA three-way junction. *Biophysical Journal* **68**, 266-274 (1995).

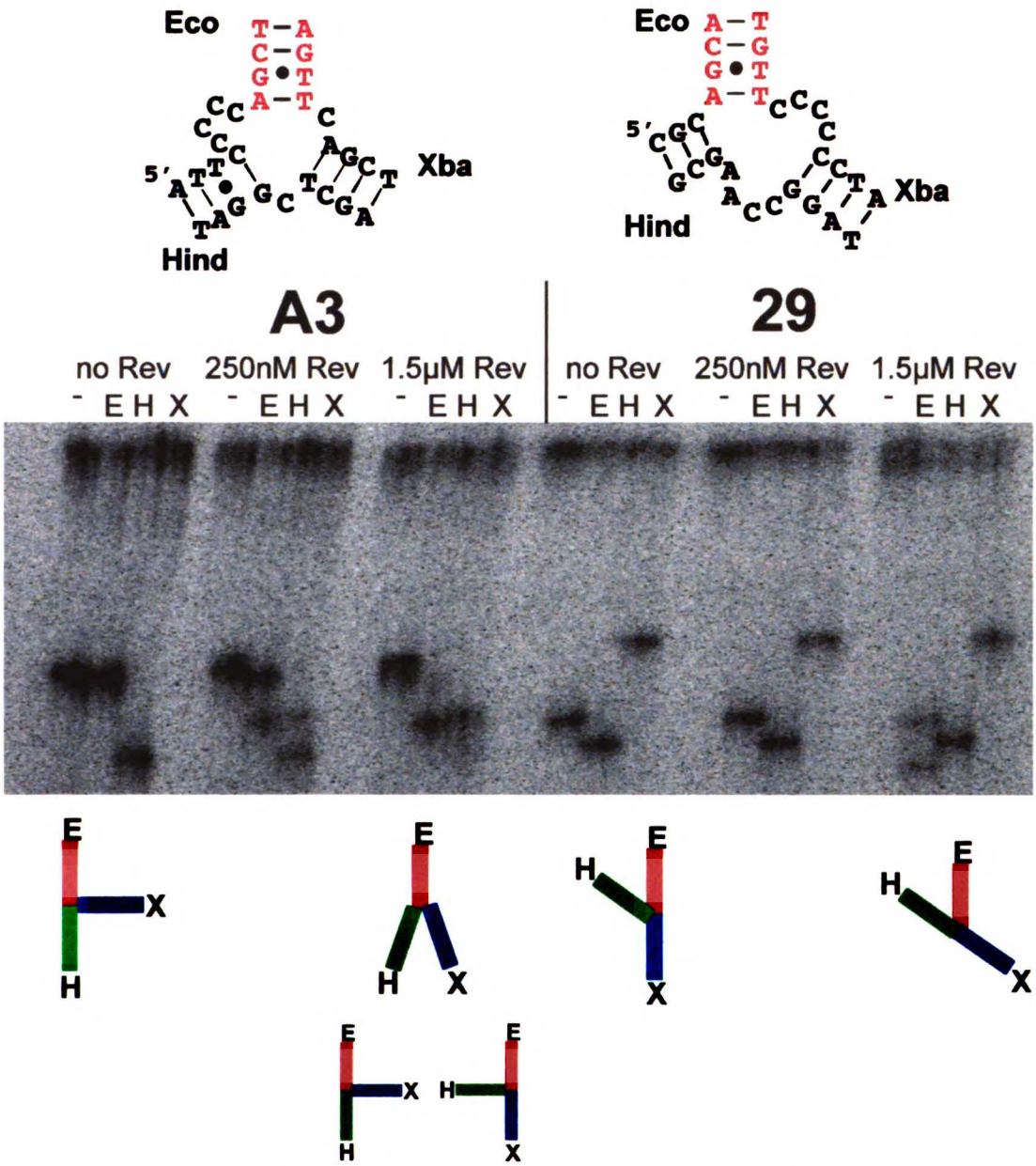


Figure 1. Conformational analysis of 3-helix junction molecules A3 and 29 Top: Core junction sequences that were incorporated into extended helix arms along with restriction site present in each arm. Middle: Representative gel showing the mobility of each molecule in the presence of the indicated amount of Rev after digestion with no enzyme (-), EcoRI (E), HindIII (H), or XbaI (X). Bottom: Interpretation of the predicted 3-helix conformations based on gel mobilities.

Appendix 4

Selection of DNA molecules that bind the BIV Tat peptide

The goal of the in vitro selection experiment described in Chapter 3 was to find DNA sequences that would illustrate some potentially general mechanisms of ssDNA tertiary structure recognition. The experiments presented in chapters 3 and 4 describe the use of the arginine-rich, α-helical peptide from HIV Rev as a target for DNA binders. Rev was chosen in part because α–helices are by far the most common units of structure used by proteins to recognize dsDNA, which suggests they might be adept at recognizing structured ssDNA. However, we also wanted to test the ability of other protein folds to bind ssDNA. A β–sheet RNA-binding structure would be ideal because this fold has also been shown to constitute a DNA recognition motif [1]. Additionally, in the case of BIV Tat, we have a well characterized arginine-rich peptide which binds structured RNA with high affinity as a β–hairpin, which should provide an ideal example to compare to the DNA-binding properties of Rev. Finally, since BTat consists of only two short β-strands, stabilized by no more that two hydrogen bonds, the structure is not stable in the absence of nucleic acid, meaning that we would actually be able to screen a variety of peptide conformations in addition to a β–hairpin for their ability to bind ssDNA[2,3].

I carried out an in vitro selection for ssDNA molecules that bound the 17 amino acid BTat peptide, using a protocol similar to that described in Chapter 3 (see detailed protocol below). However, the progression of this selection was quite different from that of the Rev selection. Whereas a clear increase in the specific affinity of the DNA pool for Rev-agarose emerged after 5 rounds of selection in 400mM NaCl, prompting an ultimate increase in the NaCl concentration to 600mM, it took 10-11 rounds of selection at 450mM NaCl to detect an increase in the amount of DNA specifically retained by BTat-agarose, and no further increase in retention could be enriched for at higher salt

concentrations (Fig. 1a,b). It is unclear why the selection proceeded at such a reduced pace, although it probably reflects a smaller number of positives in the initial population, either because BTat is an inherently weaker ssDNA binder than Rev or because these selection conditions disfavor tight binding.

It was also observed that, while the later-round DNA pools were relatively resistant to competition from a BTat peptide with all arginines mutated to lysine (BTat R->K), the Rev peptide was able to compete effectively with wild-type BTat (Fig. 1c). This may not be surprising, because while the BTat peptide has 7 arginines and 1 lysine, Rev has 11 arginines, which should contribute significantly to non-specific, electrostatically-drive affinity. To attempt to circumvent this, I incorporated a negative selection in rounds 13 and 14 in which I eliminated DNA eluted from the BTat agarose by the Rev peptide. As shown in Figure 1c, this significantly reduced the affinity of the population for Rev.

At this point, 18 molecules were cloned and sequenced. No strong sequence homology within the population was discovered, so 10 molecules were chosen and assayed for their ability to compete with the BIV TAR RNA stem-loop for binding to BTat in a gel shift assay (Fig.2a). Only one these molecules, sequence #2, was able to act as a high affinity competitor against the natural peptide:RNA interaction . Quantitation shows that this sequence is only a 2-fold weaker competitor than cold BIV TAR RNA, while the next best DNA sequence, 3-9, is at least an 8-fold weaker competitor.

I next decided to try and map the BTat binding site on sequence #2 using the methods outlined in Chapter 3 and Appendix 1. Figure 3a shows the results of the truncation analysis for this molecule. It appears that ~7 nucleotides can be removed from the 5' end of the molecule and ~22 nucleotides from the 3' end without disrupting BTat binding.

The computationally predicted stable fold that is most consistent with this is a 3-helix junction (Fig. 3c) and, significantly, it possesses a G•T/CG basestep, identical to the ones in the ssDNA binding sites selected for Rev. To determine if these features are part of the binding site, I used DMS and ENU modification interference to identify positions that are important for peptide recognition. These results, shown in Figure 3b and summarized in Figure 3c, show that the important positions are indeed clustered at or near the 3-helix junction, and include the G•T/CG.

Although my next objective was to use DNA and peptide mutants to more carefully characterize the binding determinants, the fluorescence anisotropy based assay used to measure the Rev-ssDNA interaction (Chapters 3 and 4) does not work for the DNA #2/BTat interaction. It is unclear why this is, although it is not a fundamental problem with measuring BTat by this sort of an assay, since the BTAR RNA:BTat interaction is easily detected this way. However, due to the lack of a simple binding assay, I have not characterized this interaction further.

It is interesting, however, that the G•T/CG motif also appears in this context. While it is possible that BTat does bind this DNA in an α-helical conformation, this is unlikely given that BTat has two glycine residues in the center of its sequence, which should disfavor α-helix formation. BTat, however, does have pairs of arginines spaced anywhere from 0 to 4 amino acids apart and, since the peptide is probably unstructured, there is likely enough flexibility in the peptide to adapt to the geometric requirements of a particular DNA structural feature. It is also important to note that the G•T/CG is experimentally unverified and more biochemistry will be needed to confidently assign it to this molecule. However, if it is part of the binding site, it further establishes this

basestep as a modular ssDNA sequence motif that is can allow high affinity recognition by many arginine-rich proteins.

Materials and Methods

BTat-agarose preparation

The BIV Tat peptide spanning residues 65-81 of the protein CSGPRPRGTRGKGRRIRR was synthesized with an N-terminal cysteine residue and a C-terminal amide were synthesize and HPLC purified. Peptide was coupled t o agarose beads exactly as described for Rev in the Materials and Methods section of Chapter 3, and resuspended at a concentration of 50 ng/μL.

In vitro selection

The DNA library used for the selection is the same as in the Rev selection described in Chapter 3. All binding experiments were done in 10mM Tris pH 7.5, 450mM NaCl, 5mM KCl, 5mM $CaCl_2$, 3mM $MgCl_2$ at 25°. For the initial round, 45μL BTat-agarose was incubated in 100μL with ~250 pmol library DNA for 30 min., and washed 3 times with 300μL buffer containing 5ug yeast tRNA (Invitrogen). DNA was eluted in two 150μL washes in buffer containing 10μM BTat (SGPRPRGTRGKGRRIRR-am) and 5ug yeast tRNA . Under these conditions, ~3% of library DNA was retained and ~20% eluted, compared to values of 21% retention and 88% elution for BTAR RNA. All PCR steps are exactly as described in Chapter 3.

For all subsequent rounds, 3-15 pmol DNA and 3ug tRNA were incubated with 15μL resin as above with the following changes: in rounds 12-14, elutions were done with 3.3μM BTat and were done only once; in round 12, beads were washed twice with 150μL 10μM BTat (R->K), where all arginines are mutated to lysine; in rounds 13-14, beads were washed four times with 150μL 10μM Rev (see Chapter 3). After 14 rounds DNA was amplified, digested with BamHI and SalI, cloned into pUC19, and sequenced.

Truncation analysis and modification interference

DNA was digested and prepared for truncation analysis as described in Appendix 1. DNA was chemically modified and processed as described in Chapter 3. DNA was fractionated with 5μL BTat-agarose and eluted twice in 100μL 10μM BTat. Gels were as described in Appendix 1.

Gel shift competition assay

A 28-nucleotide BIV TAR hairpin transcript was body labeled with α-CTP during in vitro transcription with T7 RNA polymerase. RNA was purified from 15% PAGE gels and precipitated. All nucleic acids were heated to 80° for 5 min. and slow cooled to 4° prior to assay. Labeled RNA and competitor nucleic acids were pre-mixed in 2x buffer (20mM Hepes pH 7.5/200mM KCl/2mM $MgCl_2$/1mM EDTA/100μg/mL yeast tRNA/20% glycerol) at 4° and an equal volume of BTat diluted in water was added. Reactions were allowed to equilibrate for 30 min. prior to separation on 10% polyacrylamide/.5xTBE gels.

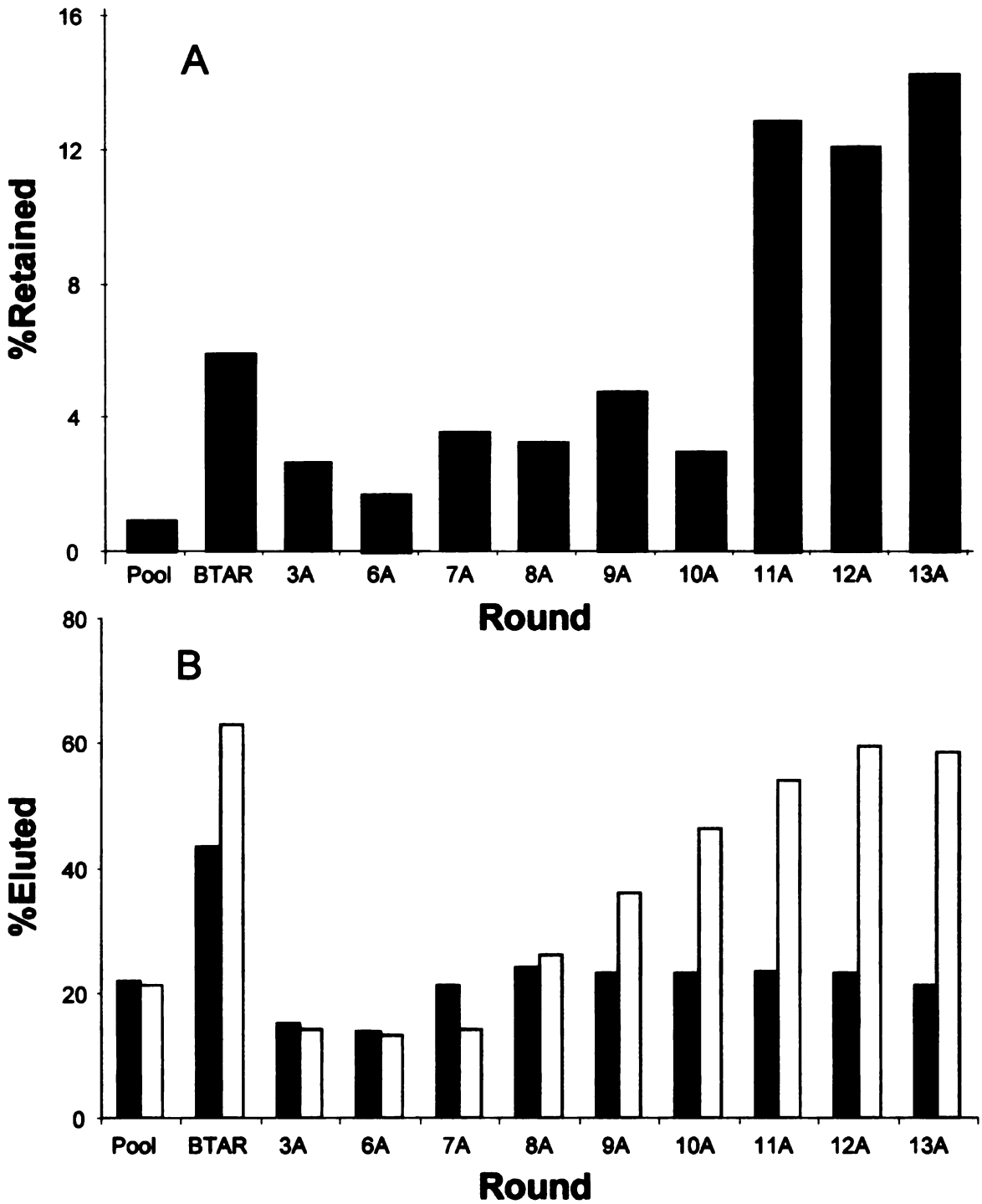
References

1. Raumann, B., Rould, M., Pabo, C. & Sauer, R. DNA recognition by beta-sheets in the Arc-repressor-operator crystal structure. *Nature* 367, 754-757 (1994).

2. Puglisi, J.D., Chen, L., Blanchard, S. & Frankel, A.D. Solution structure of a bovine immunodeficiency virus Tat-TAR peptide-RNA complex. *Science* 270, 1200-1203 (1995).

3. Ye, X., Kumar, R.A. & Patel, D.J. Molecular recognition in the bovine immunodeficiency virus Tat peptide-TAR RNA complex. *Chem. Biol.* 2, 827-840 (1995).

Figure 1


A
%Retained
16
12
8
4
0
Pool
BTAR
3A
6A
7A
8A
9A
10A
11A
12A
13A
Round
B
%Eluted
80
60
40
20
0
Pool
BTAR
3A
6A
7A
8A
9A
10A
11A
12A
13A
Round

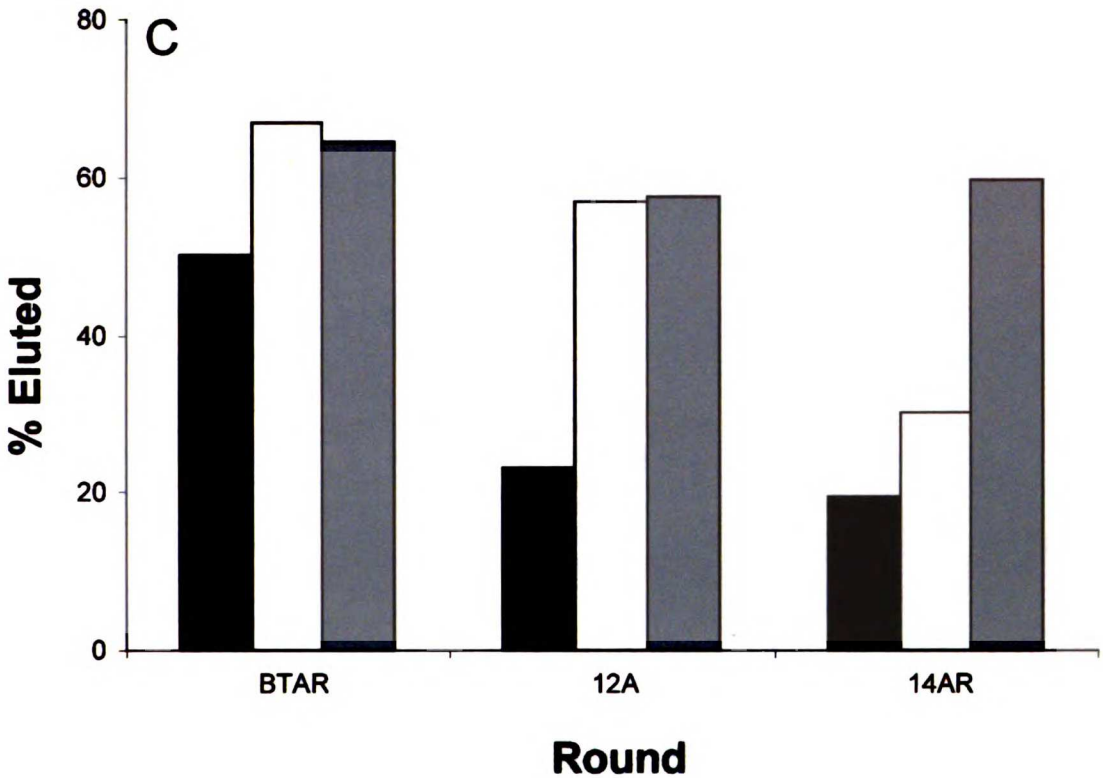
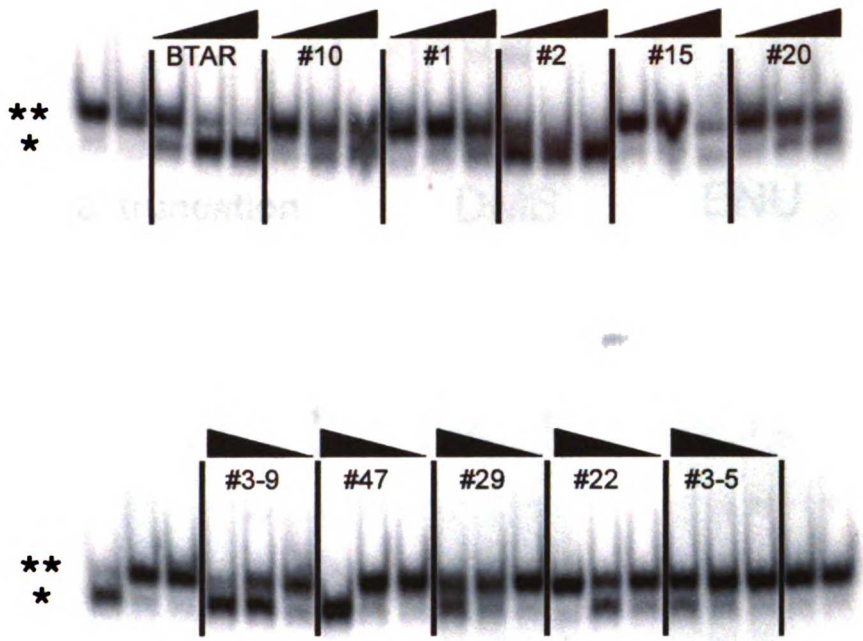


Figure 1. Progress of selection for ssDNA molecules that bind BTat. (A) Percentage of input nucleic acid retained on BTat-agarose following washes. Pool is unselected starting nucleic acid and BTAR is 28 nucleotide high affinity RNA binding site for BTat. (B) Percentage of bound nucleic acid eluted with BTat (R->K) peptide (filled bars) or BTat peptide (open bars). (C) Percentage of bound nucleic acid eluted with BTat (R->K) peptide (filled bars), Rev peptide (open bars), or BTat peptide (grey bars). DNA from Round 14AR has undergone two rounds of selection against binding to the Rev peptide (see Materials and Methods).

Figure 2

A



B

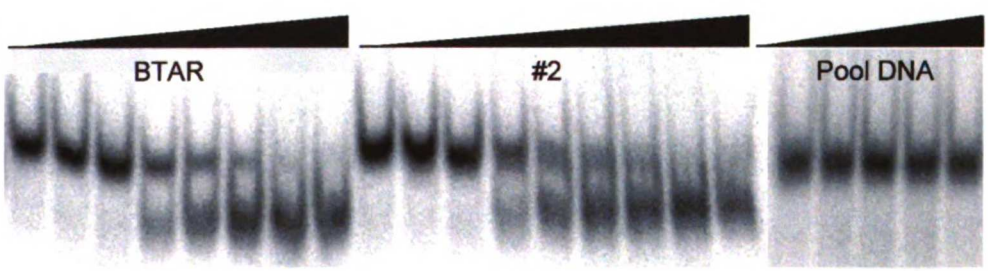
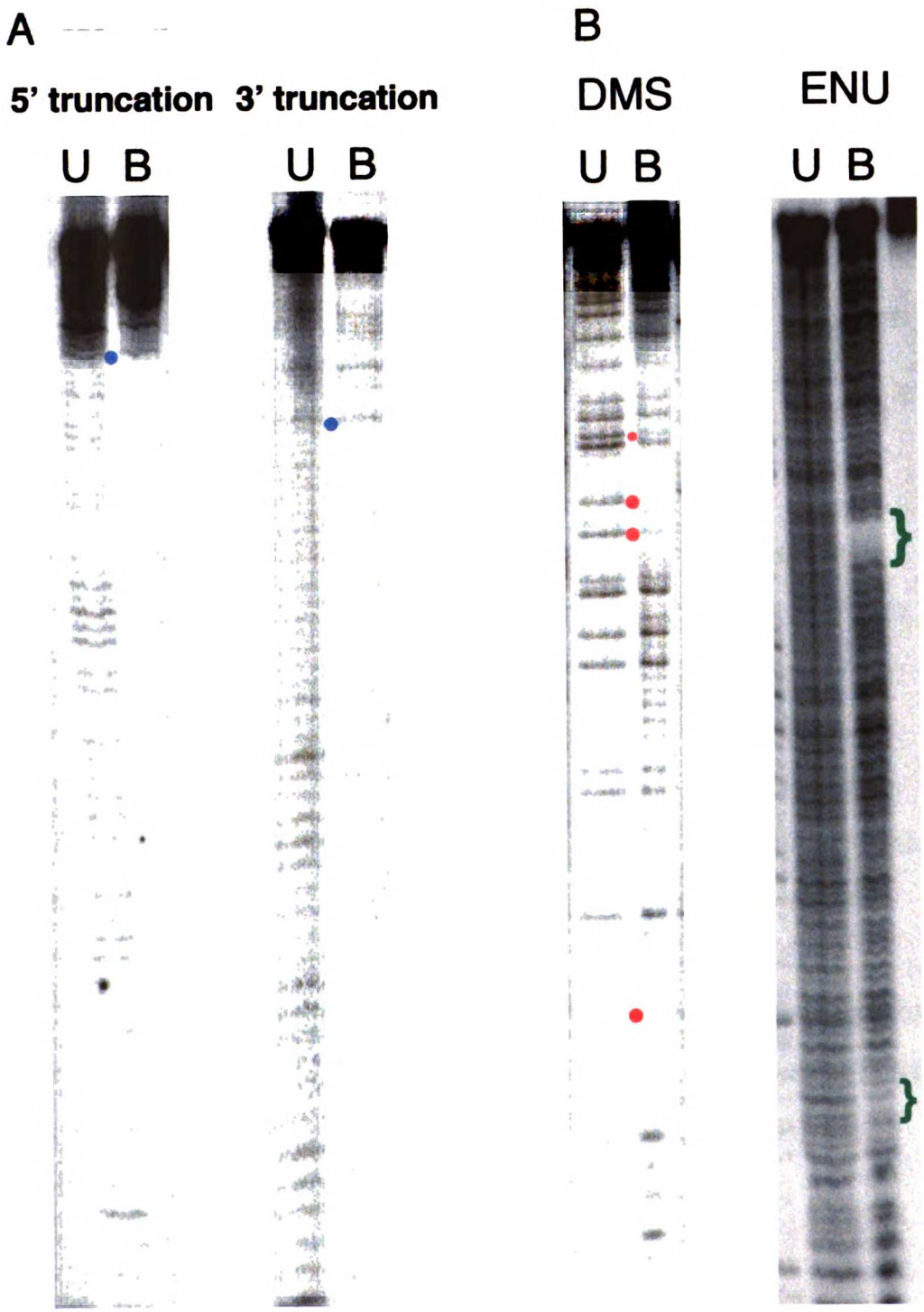


Figure 2. One selected sequences competes efficiently with BTAR for BTat binding. (A) Competition gel shifts with 2nM labeled RNA and 250nM BTat peptide. * indicates unshifted RNA, ** indicates peptide bound, shifted RNA. Names indicate the identities of the competitor sequences. Concentrations of competitors are 250, 1250, and 6250 nM. (B) Competition gel shift using same conditions as in 2a. For BTAR, competitor concentration ranges from 20-1280nM. For DNA #2, the range is 20-2560nM. For pool DNA, the range is 640-10240nM. All competitor concentrations increase in two-fold step sizes.

Figure 3



C

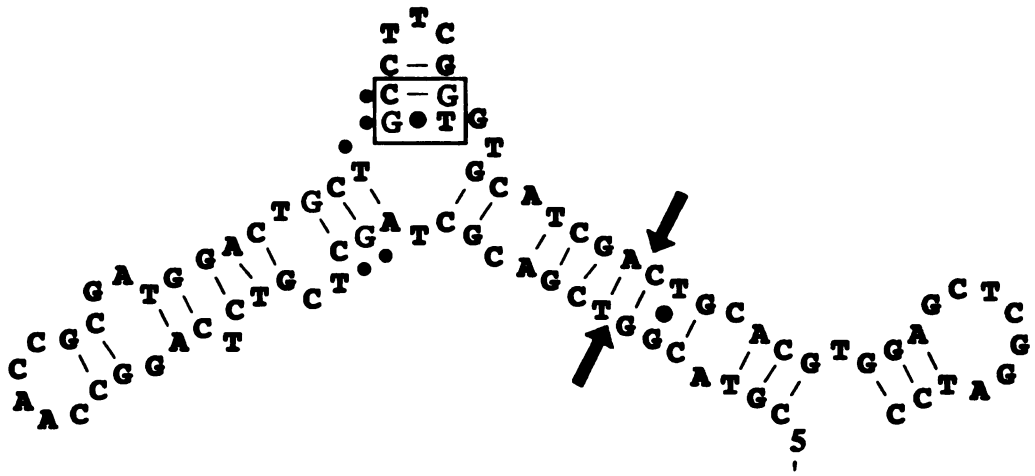


Figure 3. Location of the BTat binding site on DNA #2. (A) Truncation analysis using 3' end labeled DNA (left) or 5' end-labeled DNA #2 (right). U indicates DNA that is unfractionated, while B is DNA that bound BTat-agarose. Blue dots indicates apparent 5' and 3' boundaries. (B) Chemical modification interference using dimethylsulfate (DMS) and N-ethyl-N-nitrosourea (ENU). U and B are as in 3a. Red dots are positions where DMS modification strongly (large dots) or moderately (small dot) interferes with binding. Green brackets indicate site where ENU modification interferes with binding. (C) Summary of data from 3a and 3b on the most likely fold for BTat DNA #2. Red indicates sensitivity to DMS, green dots show sensitivity to ENU, blue arrows show 5' and 3' boundaries. Box highlights G•T/CG motif.

Chapter 5

Conclusions and Future Directions

The work described in this thesis provides evidence that the distinction frequently made between proteins that bind DNA and those that bind RNA is less meaningful than it is often taken to be. This has been suggested by the previous observations that the RNA-recognition motif (RRM) and hnRNP K-homology domain (KH) protein families both have members that function by binding DNA [1,2]. However, my work with proteins of the ARM family represents a more stringent test of this idea. These proteins make extensive contacts in the deep major groove of RNA, an environment that DNA is unlikely to mimic, and we show that, with the BTat-BTAR example, binding to DNA is far weaker than to RNA. By looking at smaller sections of the binding interface however, we found that several energetically critical contacts between peptide and RNA can either also be made between peptide and DNA or that the peptide can adapt to the DNA structure using alternate interactions. We extended this even further by showing that Rev can bind DNA molecules with an affinity and specificity that meets or exceeds that for RNA.

While the distinction between DNA-specific and RNA-specific may be problematic in describing these peptides, their classification as structured nucleic acid-specific is more useful. The naturally occurring Rev binding site is a hairpin with a purine-rich internal loop (Figure 1a, Chapter 3). All the purines in the loop are engaged in non-Watson Crick basepairing and coaxial stacking in the A-form helix is never disrupted [3,4]. The molecules we have identified as tight Rev binders are defined, in one case, by a hairpin in which only one important residue is unpaired and, in the other, by an extensively basepaired 3-helix junction with also just one conserved unpaired residue (see Figure 1d, Chapter 3). Since all the evidence suggests the DNA molecules are in the B-form, it is

clear that there is a requirement for secondary structure in the binding site, regardless of its helical geometry. Together, this indicates that ARM proteins, and probably other proteins specific for structured RNA, have the physical characteristics of good structured-DNA binding proteins.

The results of our selection suggest that this might also make surprisingly good biological sense. The G•T/CG sequence motif that we have identified as essential for tight binding of Rev to ssDNA is simple and, since the G•T basepair is the most stable of all non-Watson Crick basepairs and the G•T/CG basestep is the most stable of all G•T basesteps, it should make a reasonable contribution to the stability of heavily structured sequences[5]. The fact that it functions in the enkephalin promoter when bound by CREB, and presumably in the molecule selected to bind BIV Tat, in addition to being present three times in the two independent Rev binders also demonstrates that it is adaptable to different nucleic acid structural contexts and to recognition by different protein structures[6,7]. Finally, the protein requirements for recognition, two arginine residues, ideally spaced one turn apart in an α-helix, are certainly not difficult for many nucleic acid binding proteins to meet.

The mode of structure specificity that we have identified is also exceptionally simple, requiring only an aromatic amino acid sidechain and the accessible end of a nucleic acid helix. However, we show that it can be comparable in value energetically to recognition of the G•T/CG and that it is highly specific, especially for the aromaticity of the sidechain. These requirements are not unreasonable for ssDNA in the cell, since multi-helix junctions are a consequence of any basepaired, extruded ssDNA structure and since tryptophans, for example, are known to be important components of nucleic acid:protein

interfaces in specific cases of RNA recognition as well as being defining features of a whole family of transcription factors[8,9]. Combined, G•T/CG recognition and junction recognition comprise a mechanism of ssDNA recognition that should generally prefer lower energy target structures. For these reasons, I would argue that the frequency with which ssDNA structure are recognized in vivo by proteins is higher than would have been predicted prior to these results.

The next phases of this work will proceed in two directions, the first being structural. We have identified and quantified the critical interaction between di-arginine and the G•T/CG basestep. We have also proposed a series of different ways the arginines might pair to recognize this motif. Now, we need to use higher resolution approaches to determine which of these proposed structural possibilities is occurring or if multiple arrangements contribute to the affinity. The R28 hairpin, which has two G•T/CG repeats, oriented oppositely to one another and recognized by arginines spaced slightly differently, is the ideal model in which to address this. This complex is small and previous NMR studies have been carried out on Rev and its interaction with RRE IIB RNA, suggesting that a high-resolution structure of Rev with R28 is a very reasonable goal[4].

The recognition of the TGTTC/3-helix junction molecules raises an additional set of structural issues. In this case, we want to understand what the conformation at the junction is and how it is specified by the helices. We also want to know why three helices in particular were selected for. Our best guess is that the relationship between the three is important because it allows two to stack coaxially and relegates the third to an unstacked conformation. Determining which helices are in each position, whether these

orientations are fixed or to what extent they are flexible, and with which of these helices the tryptophan is stacking would then be prime objectives in the study of these molecules. Therefore, we are also pursuing an NMR structure of the A3 molecule described in Chapters 3 and 4. Since the question of the orientation of the 3 helices is largely a question of the orientation of flexible domains, the NMR study will also include the analysis of residual dipolar couplings (RDC). Conventional NMR techniques rely on the detection of atomic interactions that are close in space, leading to the propagation of error over many steps and uncertainties in the estimation of the relative positions of separate domains. RDC analysis is a distance-independent technique that has been applied to branched nucleic acids such as the hammerhead ribozyme and should be well suited to the analysis of these molecules[10,11].

The second direction for this work involves approaches to expanding on the implications of our in vitro selection results for ssDNA function in the cell. Although I believe that the selected molecules have many of the characteristics necessary to carry out a biological function, this is obviously speculative. Therefore, one goal will be to develop reporter-based in vivo assays to directly test for biological activities. There are several promising systems that might be adapted to this purpose. The simplest approaches will be to use the reporter assay developed by Levens and colleagues to measure the activity of hnRNP K and the related FUSE family of proteins. They have adapted a simple one-hybrid assay to look at transcriptional activation via ssDNA, measuring the ability of a fusion between hnRNP K and the Gal 4 transcriptional activation domain to activate a reporter with hnRNP K sites in the promoter [12]. Since I propose that the mechanism of recognition of the enkephalin promoter hairpin by CREB

is fundamentally similar to the mechanism by which Rev binds R28, it might also be worthwhile to use this promoter context to test the Rev-R28 interaction. Rev fusions to CREB or its activation domain would also be able to take advantage of the fact that this site is adapted to cruciform formation by its location in the promoter and its proximity to potential cooperating sites[13].

Although these approaches are promising, it would be ideal to also develop a prokaryotic reporter system. The value of this would be that the superhelical forces on DNA in bacteria can be modulated with topoisomerase mutants or with environmental stimuli such as osmotic shock or antibiotics [14,15]. Cruciform formation will usually be disfavored energetically, so having a system in which the forces that act on chromosomal DNA can be readily modified may be important. One approach to doing this would be to use the well characterized interaction between the bacteriophage N4 RNA polymerase and a DNA hairpin, located at the –12 position of the phage early promoters, by making Rev fusions to the enzyme[16,17]. Additional reporters incorporating Rev fusions to other transcriptional activators in a one-hybrid assay, similar to that proposed above, can also be imagined[18].

With reporter systems in hand, several issues can be addressed. One involves the kinetic and thermodynamic requirements for ssDNA extrusion. An ideal ssDNA site will be thermodynamically competitive with its duplex state, but it will also have to form with reasonable kinetics, which are generally dependent on the rate of initial duplex melting[14]. This is generally much greater in AT-rich sequences, while the R28 sequence is relatively GC rich in the essential regions. It will be interesting to determine, first, if R28 can be made to form and, second, if its formation can be improved by altering the

sequence context around the essential positions. This could include adding poly AT stretches around the hairpin or testing sequences from naturally extruding hairpins such as enkephalin. Another idea is to take the 3-base loop sequence from the N4 promoter hairpin, which has been shown to be critical for catalyzing and stabilizing the extruded structure, and see if it promotes extrusion when used to replace the R28 loop[19,20]. It will also be interesting to see if more complex structures such as the 3-helix junction are at all accessible from chromosomal DNA. Finally, a robust reporter system offers the opportunity to carry out library screens for enhanced functional characteristics. For instance, although Rev appears to be specific for 3-helix junctions, a cruciform-specific peptide would be a more valuable model for a potential biological interactions. Using an in vivo system, DNA structure libraries or peptide mutants could be screened to identify these sorts of specificities.

Finally, it will be interesting to see if the results presented in chapter 3 can be exploited to identify new ssDNA structures in vivo. In principle, likely candidates for ssDNA sites can be crudely predicted based on the presence of inverted repeats, which have the potential to form cruciforms, and such attempts have been published[21]. It will be very interesting to combine these predictions with a search for G•T/CG motifs. Additional predictors can also be included, such as phylogenetic conservation of structures and sequence, and the presence of particularly stable structures such as the N4 loop. These sites would likely represent a class enriched for ssDNA structures that functions via specific protein binding, and the structural models we have for predicting G•T/CG recognition would also help in identifying the protein partner. This would mark

the exciting transition from the study of model interactions described in this thesis to the study of the biological effectors we believe are mostly yet undiscovered.

References

1. Michelotti, E., Michelotti, G., Aronsohn, A. & Levens, D. Heterogeneous nuclear ribonucleoprotein K is a transcription factor. *Molecular and Cellular Biology* **16**, 2350-2360 (1996).

2. Ding, J., Hayashi, M., Zhang, Y., Manche, L., Krainer, AR & Xu, R. Crystal structure of the two-RRM domain of hnRNP A1 (UP1) complexed with single-stranded telomeric DNA. *Genes and Development* **13**, 1102-1115 (1999).

3. Ye, X., Gorin, A., Ellington, A.D. & Patel, D.J. Deep penetration of an alpha-helix into a widened RNA major groove in the HIV-1 rev peptide-RNA aptamer complex. *Nat. Struct. Biol.* **3**, 1026-1033 (1996).

4. Battiste, J.L. *et al.* Alpha helix major groove recognition in an HIV-1 Rev peptide-RRE RNA complex. *Science* **273**, 1547-1551 (1996).

5. Allawi, H. & SantaLucia, J. Thermodynamics and NMR of internal G•T mismatches in DNA. *Biochemistry* **36**, 10581-10594 (1997).

6. Spiro, C., Richards, J., Chandrasekaran, S., Brennan, R. & McMurray, C. Secondary structure creates mismatched base pairs required for high-affinity binding of cAMP response element-binding protein to the human enkephalin enhancer. *Proceedings of the National Academy of Science* **90**, 4606-4610 (1993).

7. Spiro, C., Bazett-Jones, D., Wu, X. & McMurray, C. DNA structure determines protein binding and transcriptional efficiency of the proenkephalin cAMP-responsive enhancer. *Journal of Biological Chemistry* **270**, 27702-27710 (1995).

8. Legault, P., Li, J., Mogridge, J., Kay, L.E. & Greenblatt, J. NMR structure of the bacteriophage lambda N peptide/boxB RNA complex: recognition of a GNRA fold by an arginine-rich motif. *Cell* **93**, 289-299 (1998).

9. Escalante, C., Yie, J., Thanos, D. & Aggarwal, A. Structure of IRF-1 with bound DNA reveals determinants of interferon regulation. *Nature* **391**, 103-106 (1998).

10. Lipsitz, R. & Tjandra, N. Residual dipolar coupling in NMR structure analysis. *Annual Review of Biophysics and Biomolecular Structure* **33**, 387-413 (2004).

11. Bondensgaard, K., Mollova, E. & Pardi, A. The global conformation of the hammerhead ribozyme determined using residual dipolar coupling. *Biochemistry* **41**, 11532-11542 (2002).

12. Tomonaga, T. & Levens, D. Activating transcription from single stranded DNA. *Proceedings of the National Academy of Sciences* **93**, 5830-5835 (1996).

13. Spiro, C. & McMurray, C. Switching of DNA secondary structure in proenkephalin transcriptional regulation. *Journal of Biological Chemistry* **272**, 33145-33152 (1997).

14. Dayn, A. *et al.* Formation of (dA•dT)n cruciforms in Escherichia coli cells under different environmental conditions. *Journal of Bacteriology* **173**, 2658-2664 (1991).

15. Dayn, A., Malkhosyan, S. & Mirkin, S. Transcriptionally driven cruciform formation in vivo. *Nucleic Acids Research* **20**, 5991-5997 (1992).

16. Glucksmann, M.A., Markiewicz, P., Malone, C. & Rothman, D.L. Specific sequences and a hairpin structure in the template strand are required for N4 virion RNA polymerase promoter recognition. *Cell* **70**, 491-500 (1992).

17. Glucksmann-Kuis, M.A., Dai, X., Markiewicz, P. & Rothman-Denes, L.B. E. coli SSB activates N4 virion RNA polymerase promoters by stabilizing a DNA hairpin required for promoter recognition. *Cell* **84**, 147-154 (1996).

18. Huffman, J. & Brennan, R. Prokaryotic transcription regulators: more than just the helix-turn-helix motif. *Current Opinion in Structural Biology* **12**, 98-106 (2002).

19. Dai, X., Kloster, M. & Rothman-Denes, L. Sequence-dependent extrusion of a small DNA hairpin at the N4 virion RNA polymerase promoters. *Journal of Molecular Biology* **283**, 43-58 (1998).

20. Dai, X., Greizerstein, M., Nadas-Chinni, K. & Rothman-Denes, L. Supercoil-induced extrusion of a regulatory DNA hairpin. *Proceedings of the National Academy of Sciences* **94**, 2174-2179 (1997).

21. Chen, S. & L, S. Identification of long intergenic repeat sequences associated with DNA methylation sites in Caulobacter crescentus and other α-proteobacteria. *Journal of Bacteriology* **185**, 4997-5002 (2003).

Appendix 5

A Tat-fusion system for identifying RNA binding proteins and its application to the Mason-Pfizer Monkey Virus CTE

The development of genetic assays to monitor protein-protein, DNA-protein, and RNA-protein interactions has greatly facilitated structure/function studies and cloning of interacting partners. Here, I describe a screening method I have helped develop for studying RNA-protein interactions in mammalian cells that appears to be relatively adaptable and may be particularly suitable for studying mammalian complexes that may, for example, require post-translational modification or multiple cellular components for binding. I also describe my role in our first attempts, made in collaboration with Rob Nakamura, to identify novel RNA-binding factors from a cDNA library, which I have created specifically for this application. The method has been described more thoroughly elsewhere [1], and a manuscript describing the cDNA library screen is in preparation.

Overview of the method

The method, called the Tat-fusion system, is based on transcriptional activation by HIV-1 Tat (see Chapter 2). Tat is an unusual transcription factor that operates by binding to an RNA hairpin, known as TAR and Tat also enhances transcription when delivered to the RNA via a heterologous RNA-protein interaction [2-8]. In the Tat-fusion system, a library is fused to the activation domain of Tat and RNA-binding domains are identified by their ability to activate an HIV-1 LTR reporter in which TAR is replaced by an RNA site of interest. The system appears to accommodate a wide variety of interactions, operates in different cell types, and in many cases shows high levels of activation.

The system uses two plasmids, a Tat-fusion expressor plasmid and a reporter plasmid expressing green fluorescent protein (GFP) from a modified HIV-1 LTR.

Reporter cell lines (typically HeLa) are generated containing the stably integrated reporter, and the Tat-fusion library is introduced into cells by bacterial protoplast fusion, which delivers library members to cells in a nearly clonal manner [3]. Cells expressing high levels of GFP (as a result of RNA binding and Tat activation) are isolated by fluorescence-activated cell sorting (FACS) and plasmid DNA is recovered. Plasmids are reintroduced into bacteria, protoplasts are generated, and fusion and sorting are repeated until the population is highly enriched for GFP expressors. Individual positive clones are then tested for the desired RNA-binding specificity by measuring activation of several mutant or unrelated RNA reporters. Clones that show the proper specificity are sequenced and characterized further.

Considerations for the Tat fusion and reporter plasmids

RNA-binding domains are fused to the complete first exon of Tat (following amino acid 72), which includes the nuclear localization sequence and TAR-binding domain. This will usually ensure proper localization of fusions and allows activity of the Tat portion to be assessed, as fusions to Tat_{1-72} are expected to be active on the wild-type HIV-1 LTR unless they are poorly expressed or disrupt the Tat-TAR interaction. It is possible that retaining the TAR-binding domain will cause nonspecific binding to some RNA reporters, but this has not yet been observed.

At least two other difficulties may be envisioned for the Tat-fusion system. First, large domains fused to Tat might sterically hinder the formation of transcription elongation complexes. Preliminary experiments with cDNA libraries, however, indicate that a substantial fraction of large fusion proteins still activate through HIV-1 TAR.

Second, it is possible that endogenous nuclear proteins or RNAs might compete for RNA binding and thereby prevent activation. However, this does not appear to have been a problem for one case we have tested, the U1A-U1 interaction (the protein and RNA are highly abundant nuclear components) and, in fact, it generally has been difficult to use untethered RNA-binding domains as dominant negative inhibitors of Tat. This likely reflects a requirement for multiple interaction surfaces to achieve Tat activation.

The reporter plasmid contains the HIV-1 LTR, followed by a polylinker for cloning RNA target sites in place of TAR, followed by an IRES (internal ribosome entry site) and then the GFP gene. An IRES was included because structured RNA elements, or the binding of proteins to these elements, can block cap-mediated translation when located near the 5' end of a transcript[9]. By initiating translation internally, it also is possible to clone RNAs containing AUG codons upstream of GFP, which might produce an out-of-frame translation product and induce termination at premature stop codons if cap dependent translation was relied upon. The vector also contains the neomycin resistance gene for selecting stable reporter-containing cell lines using G418.

We typically include the BIV TAR hairpin (see Chapter 2) immediately downstream of the RNA target site for two reasons. First, because Tat activates transcription through an elongation mechanism, it is possible that RNAs beyond a certain length are not synthesized before transcription terminates and thus Tat would not be recruited to the transcript [10]. By placing BIV TAR downstream of the target site, transcription past the site can be monitored by measuring activation through BIV TAR using the HIV-BIV Tat_{65-81} fusion protein to deliver the HIV-1 Tat activation domain[5]. We have observed levels of GFP activation that are readily detectable by FACS when

BIV TAR is located more than 170 nucleotides from the 5' end. Thus, it appears that RNAs at least 100-200 nucleotides in length can be accommodated by the Tat-fusion system. Second, BIV TAR is useful for identifying appropriate reporter cell lines for screening libraries. Individual clones are transfected with the HIV-BIV Tat_{65-81} protein to test whether the reporter can be activated, and clones showing the highest levels of activation are chosen for screening.

cDNA library screening

A cDNA library screen was designed to identify proteins that interact with the Constitutive Transport Element (CTE) RNA from the Mason-Pfizer Monkey Virus. The CTE is a 160-nucleotide RNA structure that confers export on unspliced RNA from the nucleus through its interaction with cellular export factors [11]. A Tat-hybrid reporter plasmid was constructed with the CTE RNA element followed by BIV TAR and a GFP reporter. The reporter construct was integrated into HeLa cells and a cell line that displayed strong GFP signal when transfected with the BIV Tat control was selected. The cell line was also shown to be responsive to a Tat protein fused to Tap, a protein known to bind preferentially to the CTE [12]. Additionally, a cDNA library was constructed from HeLa mRNA. 6.7×10^6 independent library clones were obtained. >75% had an insert and the average insert was ~1.8kB.

Next, a large-scale screen to identify CTE-interacting clones was initiated. In five separate first round screens, over 72 million HeLa cells were analyzed, with 204,000 GFP positive cells sorted, and ~118,000 *E. coli* transformants recovered. Based on an estimated fusion efficiency of 10%, approximately 7.5 million individual clones were

analyzed in this round. The transformants were pooled, amplified, and prepared for a second round of screening. In this round, a higher GFP window was utilized to enrich for strong activators and a more conservative lower window was also set to retain weaker GFP positive clones for a possible third round of enrichment. From the high sorting window, 3.6 million HeLa cells were analyzed, 16,500 positive cells sorted, and ~5000 transformants recovered.

Following the second round library sort, we proceeded to analyze individual clones obtained from the high GFP sorting window. DNA from individual clones was prepared and digested to verify the integrity of the plasmid and insert. Approximately half of the recovered clones were discarded because the plasmids had extremely small or no inserts or contained abnormalities in the plasmid backbone. Of the remaining clones, several hundred were re-tested on the CTE reporter cell line by protoplast fusion or transfection followed by flow cytometry analysis. Clones conferring strong GFP activity were then sequenced and further analyzed.

Sequence analysis of activating clones

The strongly positive clones can be separated into three classes: RNA binding proteins that contain RRM domains and/or RGG boxes, proteins that do not contain obvious RNA binding domains, and non-coding sequences that fortuitously encoded peptides that confer reporter activity. The largest number of clones identified fell into the first category, and it is these clones that produced the strongest reporter activity. An example of this type of clone is hnRNP A1, an RNA binding protein that has been implicated in RNA export[13,14]. HnRNP A1, which contains two RRM domains, an

RGG box, and the M9 nuclear localization/nuclear export signal, was obtained multiple times in the screen in various truncated and splice variant forms[15]. Other RRM and RGG box containing proteins identified included other members of the hnRNP family, as well as hnRNP related proteins including hnRNP R, nucleolin, and the Ewing Sarcoma protein (EWS).

The second class of clones, proteins which by sequence analysis did not encode known RNA-binding domains, were not studied further. Hybrid based genetic assays such as the Tat-hybrid system can identify positives which might activate by mechanisms other than specific RNA binding and likely are not involved with CTE biology. However, these clones were clearly active on the reporter and we do not rule out the possibility that they may contain novel RNA-binding domains or activate via bridging RNA-binding proteins. The third class of proteins, out-of-frame fusions to non-coding sequences, were also not extensively studied. However, we have observed that some of the amino acid sequences in this class of proteins was biased towards R, RG, and RGG sequences. This suggested that these peptides were in fact enriched for these sequences, which are often found in RNA binding proteins in general and are further enriched in proteins selected here for CTE-association[16]. This may suggest, therefore, that RGG-like motifs are important determinants of many of the CTE interactions we have selected for.

Individual assays of CTE-interacting clones

The clones that re-tested as moderately to strongly positive for reporter activity were analyzed quantitatively using the Tat-hybrid system. In place of the GFP reporter, a CAT

reporter system was used in order to obtain a semi-quantitative measure of reporter activity. To ensure that the expression of the library clones was consistent, the clones were independently co-transfected to HeLa cells with a pHIV LTR TAR-CAT reporter and CAT activity was shown to be consistent for all clones.

We next asked whether the binding of the clones was specific to the CTE. To assess specificity, we compared activity of the clones on an unrelated RNA sequence, the TAR loop from BIV. The clones were co-transfected with a pHIV LTR BTAR CAT reporter and reporter activity was measured and compared to the results with the pHIV LTR CTE BTAR CAT reporter. Because both reporters could be activated by the HIV-BIV Tat_{65-81} fusion protein, reporter activities were normalized to the reporter signals conferred by this protein. While Tap displayed the highest specificity for the CTE, several of the library clones from the hnRNP family also bound the CTE with a high specificity.

We chose to focus further studies on hnRNP A1 for the following reasons. This protein was identified multiple times in our screen and showed a selectivity for the CTE that was nearly as high as the Tap control. HnRNP A1 has been long implicated in the export of RNA thus it was a plausible candidate for a CTE interacting protein[13,14]. Furthermore, because the architecture of hnRNP A1, an RRM containing protein with an RGG domain, is similar to that of many of the positive clones in the library, we take it as a representative of many of the clones we have selected. Finally, several N-terminal truncations of hnRNP A1 were also identified in the screen and provided a convenient means to map the CTE interaction domains of the protein.

HnRNP A1, the hnRNP A1 truncation mutants identified from the library screen, and several genetically constructed hnRNP A1 deletion mutants were used in semi-quantitative in vivo binding studies to assess the relative contributions of the various domains of the protein to the binding and specificity for the CTE. Full length hnRNP A1 was very active on the CTE reporter and was highly specific for the CTE. Two of the clones identified from the library screen, hnRNP A1 144-319 and 159-319, contained deletions of the entire first RRM and a full third or half of the second RRM domain respectively. While both the activity and specificity of these clones on the CTE reporter was lower than those of the wild-type protein, these clones both had activity and specificity for the CTE RNA. This suggested that the major RNA binding components of this protein, the RRMs 1 and 2, are partly but not entirely responsible for the RNA binding specificity, and that the RGG domain or the C-terminus of the proteins likely play a role in the specificity for the CTE.

Tat-fusions to hnRNP A1 1-194 (containing only RRMs 1 and 2) and 195-319 (containing just the RGG domain and M9) were also constructed. 1-194 was modestly active and specific for the CTE reporter, suggesting that this domain does make a contribution to CTE binding. 195-319 also showed modest activity and selectivity, consistent with the results of the 144-319 and 159-319 mutants. The activity of 195-319 on the CTE reporter clone was not as high as the other clones tested in this experiment, possibly because folding of the C-terminus of hnRNP A1 is destabilized in the absence of the RRMs. The Tat-hnRNP A1 195-319 clone protein did confer the expected level of activation on a HIV TAR reporter confirming that the transfection and expression of the protein was normal. Efforts to characterize a Tat-fusion to the RGG domain alone were

unsuccessful, consistent with the hypothesis that this region of the protein is destabilized in the absence of the major RNA binding domains.

Conclusion

The results of the library screen and subsequent analysis of library clones provides several lines of evidence that the interaction of many of the library clones with the CTE is dependent upon RGG domains. HnRNP A1, one of the strongest and most specific CTE binders, contains 4 RGG boxes and deletion analysis showed that the RGG domain was responsible for some of the activity of this protein. Many other cDNAs from the library screen also encoded proteins that contain RGG boxes and this feature was statistically enriched in the population of positive clones.

This screen also demonstrates some of the promise of this technique for identifying novel RNA-binding specificities. We are somewhat disappointed that we have cloned mostly very abundant RNA-binding proteins such as hnRNP A1, but this is likely because of the ability of the ubiquitous RGG domain to form a complex with the CTE that may also involve protein:protein interactions with Tap. In the future, built in specificity controls using a second reporter and the use of a normalized cDNA library to eliminate biases towards proteins such as hnRNP A1 will make this system more effective.

References

1. Landt, S., Tan, R. & Frankel, A. Screening RNA-binding libraries using Tat-fusion system in mammalian cells. *Methods in Enzymology* **318**, 350-363 (2000).

2. Tan, R., Chen, L., Buettner, J.A., Hudson, D. & Frankel, A.D. RNA recognition by an isolated α helix. *Cell* **73**, 1031-1040 (1993).

3. Tan, R. & Frankel, A.D. A novel glutamine-RNA interaction identified by screening libraries in mammalian cells. *Proc. Natl. Acad. Sci. USA* **95**, 4247-4252 (1998).

4. Madore, S.J. & Cullen, B.R. Genetic analysis of the cofactor requirement for human immunodeficiency virus type 1 Tat function. *J. Virol.* **67**, 3703-3711 (1993).

5. Chen, L. & Frankel, A.D. An RNA-binding peptide from bovine immunodeficiency virus Tat protein recognizes an unusual RNA structure. *Biochemistry* **33**, 2708-2715 (1994).

6. Blair, W.S. *et al.* Utilization of a mammalian cell-based RNA binding assay to characterize the RNA binding properties of picornavirus 3C proteinases. *RNA* **4**, 215-225 (1998).

7. Selby, M.J. & Peterlin, B.M. Trans-activation by HIV-1 Tat via a heterologous RNA binding protein. *Cell* **62**, 769-776 (1990).

8. Southgate, C., Zapp, M.L. & Green, M.R. Activation of transcription by HIV-1 Tat protein tethered to nascent RNA through another protein. *Nature* **345**, 640-642 (1990).

9. Stripecke, R., Oliveira, C.C., McCarthy, J.E. & Hentze, M.W. Proteins binding to 5' untranslated region sites: a general mechanism for translational regulation of mRNAs in human and yeast cells. *Mol. Cell. Biol.* **14**, 5898-5909 (1994).

10. Selby, M.J., Bain, E.S., Luciw, P.A. & Peterlin, B.M. Structure, sequence, and position of the stem-loop in tar determine transcriptional elongation by tat through the HIV-1 long terminal repeat. *Genes Dev.* **3**, 547-558 (1989).

11. Saavedra, C., Felber, B. & Izaurralde, E. The simian retrovirus-1 constitutive transport element, unlike the HIV-1 RRE, uses factors required for cellular mRNA export. *Curr. Biol.* **7**, 619-628 (1997).

12. Grüter, P. *et al.* TAP, the human homolog of Mex67p, mediates CTE-dependent RNA export from the nucleus. *Mol. Cell* **1**, 649-659 (1998).

13. Izaurralde, E. *et al.* A role for the M9 transport signal of hnRNP A1 in mRNA nuclear export. *Journal of Cell Biology* **137**, 27-35 (1997).

14. Izaurralde, E. & Adam, S. Transport of macromolecules between the nucleus and the cytoplasm. *RNA* **4**, 351-364 (1998).

15. Siomi, H. & Dreyfuss, G. A nuclear localization domain in the hnRNP A1 protein. *Journal of Cell Biology* **129**, 551-560 (1995).

16. Burd, C.G. & Dreyfuss, G. Conserved structures and diversity of functions of RNA-binding proteins. *Science* **265**, 615-621 (1994).