Chandreyee Das
Alan D. Frankel
Department of Biochemistry and Biophysics, 600 16th Street University of California, San Francisco, CA 94143-2280

Received 4 March 2003; accepted 4 March 2003

Sequence and Structure Space of RNA-Binding Peptides

Abstract: *Studies of RNA-binding peptides, and recent combinatorial library experiments in particular, have demonstrated that diverse peptide sequences and structures can be used to recognize specific RNA sites. The identification of large numbers of sequences capable of binding to a particular site has provided extensive phylogenetic information used to deduce basic principles of recognition. The high frequency at which RNA-binding peptides are found in large sequence libraries suggests plausible routes to evolve sequence-specific binders, facilitating the design of new binding molecules and perhaps reflecting characteristics of natural evolution.* © 2003 Wiley Periodicals, Inc. Biopolymers 70: 80–85, 2003

Keywords: *RNA-binding peptides; combinatorial library; sequence specificity*

INTRODUCTION

The study of peptide–RNA complexes has been particularly enlightening to those interested in molecular recognition of RNA. A number of peptide model systems have been found to accurately reflect the binding behavior of biologically important RNA-binding proteins, yet are small and simple enough to allow extensive dissection of the intermolecular interactions and to provide insight into designing RNA-binding small molecules or drugs. Several recent reviews have defined RNA-binding peptides as short (<40 amino acids) polypeptides that often adopt bound structures distinct from their free conformations, generally undergoing disorder → order transitions upon binding.[1–3] Given the sequence simplicity of many of these peptides, it is of interest to explore the amino acid "sequence space" required to define a specific RNA binder. The dimensions of this sequence space appear quite large, made apparent over the past few years by the identification of many new RNA-binding peptides, largely from combinatorial library experiments. The recognition strategies and types of structures involved also appear quite diverse, as revealed by the increasing number of structures of peptide–RNA complexes. Here we examine the boundaries of RNA-binding peptide sequence space, attempting to define characteristics that may aid in the drug design process or help establish principles of RNA recognition.

An early indication that a large number of peptide sequences might be capable of recognizing a given RNA site came from a study of methionyl–tRNA synthetase.[4] Extensive Ala and Ser substitutions were introduced at several positions of a critical RNA-binding loop and those able to function as synthetases in vivo were isolated. Nearly 200,000 amino acid sequences, or ~1% of the library (~1.9×10^7 sequences), could recognize the tRNA anticodon loop. Thus, it seemed clear that the particular peptide sequence chosen by nature was not unique, and that

Correspondence to: Alan Frankel; email: frankel@cgl.ucsf.edu
Biopolymers, Vol. 70, 80–85 (2003)

Table I Secondary Structures of Some RNA Binding Peptides

RNA Target	Peptide	Structure	Ref.
λ BoxB	λ N	Bent α-helix	8
	Nun	Bent α-helix	11
	λ N (1–36)	Bent α-helix	12
P22 boxB	P22 N	Bent α-helix	10
ϕ21 boxB	ϕ21 N	α-helix	10a
HIV-1 RRE	R_6QR_7	α-helix	16
	Rev	α-helix	14
	SFR1	Zinc finger-stabilized helix	24
	ZF2-Rev	Zinc finger-stabilized helix	19
	RSG-1.2	Extended-turn-helix	25–27
Rev aptamer I	Rev	α-Helix	15
5S RNA	TFIIIA ZF4	Zinc finger	22,23
Yeast $tRNA^{Phe}$ anticodon	T^F2	Possibly β structure	30
HIV-1 TAR	(L)Lys(D)Lys(L)Asn	Unnatural tripeptide	40
BIV TAR	BIV Tat	β-Hairpin	33,34
	JDV Tat	β-Hairpin	35
	Cyclic BIV Tat	Cyclopeptide	38
Rev aptamer II	Rev	Extended	36
Rex aptamer	Rex	S-shaped	32

RNA-binding sequences in general might be highly degenerate, perhaps facilitating the early evolution of RNA-binding proteins from an "RNA world." Extensive combinatorial library experiments summarized here are consistent with the view that many peptide sequences are compatible with recognition of a given RNA site, at least for some structured RNAs. We examine this sequence space in the context of peptide secondary structure and the known structures of RNA–peptide complexes (see Table I).

α-HELICES

The boxB hairpin of bacteriophage λ has been the target of extensive mRNA–peptide fusion library experiments, some utilizing libraries of $>10^{13}$ peptide sequences.[5–7] BoxB is normally bound by an N-terminal arginine-rich peptide of the λ N antiterminator protein, in which the peptide binds as a bent α-helix to the widened major groove of an A-form helix and adjacent 5-nucleotide loop.[8] A large number of boxB binding peptides were found from libraries in which 10 amino acids of the λ N peptide were randomized,[5–7] and their sequences were generally enriched in helix-forming amino acids and showed covariation patterns that appeared to preserve or stabilize the helix. Specific amino acids were required at only four positions. Like some of the HIV-1 Rev response element (RRE)-binding peptides described below, the boxB binders appear to present base-specific residues on one broad face of a positively charged α-helix.

Other experiments by Roberts and co-workers[9] demonstrate how sequence information can be used to design a specific RNA-binding peptide. The boxB hairpins from phages λ and P22 are both capped by stable tetraloops with an extra extruded base, and are recognized with high selectivity by their cognate N peptides.[8,10] Austin et al.[9] found that substituting three residues in the λ peptide with corresponding residues from P22 enhanced RNA-binding affinity by ~100-fold yet retained specificity for the λ loop. One substitution added a positive charge, consistent with observed sequence preferences from the selection experiments described above, and resulted in an increased salt dependence of binding. The two other substitutions did not affect the electrostatic component but rather may influence hydrophobic or van der Waals interactions, such as the stacking of aromatic amino acids with loop nucleotide bases.[8,10]

The NMR structures of additional boxB–peptide complexes further underscore the diversity of peptide sequence and recognition properties. A ϕ21 N peptide binds to its cognate boxB as a straight, rather than bent, α-helix yet recognizes its RNA loop, with differs from the other boxB sequences, in a manner that mimics the other N complexes.[10a] An arginine-rich peptide from the phage HK022 Nun protein, which competes with λ N for binding to boxB and causes transcription termination rather than antitermination,

binds RNA in a bent α-helical conformation and with an affinity similar to N.[11] However, the Nun helix is more amphipathic, creating a solvent-exposed hydrophobic surface upon RNA binding that probably mediates a protein–protein interaction. It seems evident that the solvent-exposed face of the helix is relatively free to evolve, even between polar and nonpolar surfaces, whereas the RNA-binding face largely maintains polar amino acids with high helical propensity, characteristics required for boxB recognition. Another recent NMR structure of a λ N peptide–boxB complex examined the role of residues C-terminal to the arginine-rich motif.[12] The structure of a peptide encompassing amino acids 1–36 showed that only the first 18 residues, corresponding to the arginine-rich motif, were involved in RNA binding, confirming the modular nature of the RNA-binding domain and further validating the analysis of sequences and structures derived from these short peptides.

Like boxB recognition, binding of an HIV-1 Rev peptide to the high-affinity IIB hairpin of the RRE or to a related aptamer involves an α-helical arginine-rich domain, with about two-thirds of the peptide surface in contact with the RNA.[13–15] One combinatorial library experiment performed in a mammalian cell transcription reporter assay used an arginine-rich peptide library (~20,000 sequences), largely constrained to α-helical conformations, to isolate peptides that bound the RRE with even higher affinities than Rev.[16] The consensus sequence revealed that a single Gln positioned within a polyarginine helix was likely to hydrogen bond to a G · A base pair in the RNA, replacing an Asn interaction that occurs in the Rev complex. In a larger library experiment utilizing a bacterial antitermination reporter assay, helical binders related to the Rev peptide sequence were identified,[17] and in an even larger library experiment utilizing a high-throughput kanamycin reporter version of the antitermination assay, even higher affinity RRE binders, also likely to be helical, were found.[18] In the latter experiments, the first 10 residues of a 19-amino acid peptide were randomized with charged, polar, alanine, proline, or glycine residues, followed by five fixed arginine residues and then four alanines to stabilize a helical conformation. Many of the highest affinity binders contained acidic residues that could form potential salt bridges with arginine residues and thus further stabilize a helical conformation, as observed with the boxB-binding peptides described above. Interestingly, most of these peptides also contained a single Gln residue, positioned similarly to the Gln residue found in the previous mammalian cell-based polyarginine library experiment,[16] again suggesting the importance of a Gln–G · A interaction. The high success rate in identifying high-affinity boxB- and RRE-binding arginine-rich α-helical peptides from large combinatorial libraries underscores the utility of the sequence and structural motif for designing novel RNA binders.

ZINC FINGERS

Given the importance of α-helices in RNA recognition, investigators have been exploring the use of zinc finger peptides to constrain the helix by metal coordination, thereby reducing the entropic cost of helix folding upon RNA binding. Indeed, engineering the α-helix of HIV-1 Rev into a zinc finger framework stabilized the helix and resulted in zinc-dependent RRE–RNA binding.[19] Several types of combinatorial library experiments have provided some indication of the sequence diversity of RNA-binding zinc fingers. In one study, a central zinc finger helix was randomized within the context of the three-finger, Zif268 DNA-binding protein, and phage display was used to isolate binders to an RNA triplet that was flanked on each side by DNA helices to bind the surrounding Zif268 fingers.[20] Zinc fingers could only be found to recognize RNA triplets containing G · A or C · A mismatches, consistent with the hypothesis that an α-helix cannot fit into the major groove of an A-form helix without nearby bulges or mismatches to widen the groove.[21] One dominant sequence was found, with four of the nine randomized residues in the α-helix conserving the Zif268 sequence, and four of the remaining five residues being positively charged, suggesting that electrostatic interactions were highly enriched in this selection experiment.

Darby and co-workers used phage display to isolate zinc fingers that recognize 5S rRNA or the RRE.[22–24] In an initial experiment with four randomized positions on one zinc finger helix of the nine-finger TFIIIA protein,[22] 24 peptides were found to bind specifically the 5S rRNA, all preserving Lys at position 1 and a majority preserving Ser at position 2. Significantly more diversity was observed at the remaining two positions, with a preference for hydrophobic amino acids. Next, ten positions on each of two helices of a two-zinc finger fragment of TFIIIA were randomized, with the library limited to polar and charged residues having some helical propensity and with a sequence complexity of $\sim 10^{18}$.[23] Just two sequences were found to bind 5S rRNA, neither showing homology to TFIIIA residues but again underscoring the diversity of sequence solutions to the RNA-binding problem.

The same large, two-zinc finger library was used to identify ten peptides that specifically bound to the IIB hairpin of the HIV-1 RRE.[23] Because only a small fraction of the total library was screened ($\sim 10^8$ phage), these binders were then further randomized by DNA shuffling, resulting in six peptides that bound with ~2-fold higher affinities. Interestingly, the helix of the second zinc finger showed some sequence similarities to the Rev helix, including enrichment for arginines and the presence of an asparagine. The helix of the first zinc finger was enriched in uncharged polar and acidic residues. It was shown subsequently that the arginine-rich zinc finger alone was sufficient for RRE IIB binding, but that binding affinity actually decreased when the IIB hairpin was placed in the context of the entire RRE.[24] The N-terminal zinc finger made a small contribution to binding affinity and specificity, perhaps resulting from the selected acidic residues within its helix.

NONHELICAL BINDING PEPTIDES

While α-helices provide good modules to fit the grooves of nucleic acid helices, it is evident that other binding solutions are possible and can be even more favorable. For example, another combinatorial experiment performed with the RRE IIB hairpin identified a very different class of binders. In this case, nine residues were randomized using only Arg, Ser, or Gly (R, S, G) codons in the context of flanking arginines, and high affinity binders were found using the antitermination reporter assay.[17] These peptides did not contain Asn or Gln residues previously seen in the helical RRE binders, and contained Gly residues that disfavored helix formation, consistent with CD experiments. Subsequent mutagenesis and reselection for tighter binders yielded the RSG-1.2 peptide, which introduced an acidic residue and a Pro residue that actually stabilized some helix content.[25] The selected RRE binders preserved the arginine-rich character of Rev but otherwise showed substantial sequence diversity that resulted in higher affinity interactions than with Rev itself.

The NMR structure of RSG-1.2 bound to the RRE IIB hairpin confirmed that the RNA is able to accommodate a nonhelical binding motif.[26,27] The peptide contains an extended region followed by a turn and a short α-helix. The short helix crosses the major groove, unlike most helices that bind along the groove, whereas the extended N-terminus lies along the groove. The RNA contains G · G and G · A base pairs as in the Rev-bound form but the amino acid–RNA contacts are quite different, suggesting that the RNA structure serves mainly to shape the binding site rather than to provide specific contact points. In the Rev complex, there are many hydrogen bonds to arginines and polar side chains whereas in the RSG-1.2 complex, van der Waals interactions involving Ala and Pro likely contribute to the 15-fold higher binding specificity of RSG-1.2. Given the diverse sequences and RRE-binding modes of Rev, RSG-1.2, zinc fingers, and Gln-containing peptides, it seems likely that the sequence space explored by these limited combinatorial libraries provides only a small glimpse into the types of binding motifs possible for the RRE site.

Most of the RNA binders described above involve rather arginine-rich peptides, but it seems likely that tight binders based on less highly charged sequences are possible. Indeed, in vitro selection experiments demonstrate that RNAs can form binding pockets exclusively for hydrophobic residues.[28] A phage display library experiment with a fully randomized 15-amino acid peptide library ($\sim 3\times 10^{23}$ sequences) identified U1 snRNA binders that were largely hydrophobic and generally lacked positively charged residues.[29] Peptides from the same library selected to bind the anticodon loop of yeast phenylalanyl tRNA also identified binders largely composed of hydrophobic residues, perhaps reflecting recognition of the triply methylated state of the anticodon.[29–31] Interestingly, two of these anticodon binders displayed β structure even in the absence of RNA. Thus, it seems likely that some RNA–peptide interactions will be driven largely by hydrophobic interactions, particularly in the context of larger RNAs that can fold to form discrete binding pockets, further expanding the available sequence space for RNA binders.

The NMR structures of other peptide–RNA complexes further underscore the structural diversity of RNA-binding peptides and their corresponding sequence diversity. The complex of an arginine-rich peptide from HTLV-I Rex bound to an RNA aptamer adopts an "S-shaped" fold in which extended stretches of peptide are punctuated by two turns, ending in a very short 3_{10}-helix.[32] The peptide utilizes Pro to form a turn, as does the RSG-1.2 peptide described above, and fits into an unusual "L-shaped" RNA structure formed by three helices that meet at a two-bulge junction. In another example, arginine-rich peptides from the BIV and JDV Tat proteins bind to TAR RNA as β-hairpins, in these cases utilizing Gly residues to help form β-turns and to allow the peptides to sit deeply in the RNA major groove.[33–35] It is interesting that the JDV Tat peptide also can bind to HIV-1 TAR in an apparently extended conformation, utilizing different side chains to make specific contacts to

the two RNAs.[35] This "chameleon-like" behavior also is observed with the arginine-rich domain of Rev, which binds to the RRE as an α-helix and to an RNA aptamer in an extended conformation.[36] Thus, the sequence space of RNA-binding peptides encompasses substantial conformational diversity and flexibility in addition to diverse specific side-chain contacts.

NON-NATURAL PEPTIDES

Given that RNA-binding peptides generally do not adopt stable conformations on their own, attempts have been made to stabilize the folded state using modified synthetic peptides. Cyclic versions of an HIV-1 Tat arginine-rich peptide were generated utilizing methylene linkers of different lengths, and one peptide, screened initially for nuclear localization, was able to bind HIV-1 TAR but also bound RRE RNA, perhaps as a result of the cyclic constraint.[37] In another study, a disulfide-linked cyclic BIV Tat β-hairpin peptide bound BIV TAR with 3-fold higher affinity than the noncyclized version,[38] consistent with the notion that prestabilizing peptide structure can enhance binding affinity, as described above for α-helical peptides.

In addition to the sequence space available to natural amino acids, it also has been possible to identify RNA ligands using non-natural amino acids. HIV-1 TAR binders have been identified from tripeptide libraries composed of both D- and L-amino acids and 15 possible side chains.[39] The two highest affinity peptides showed a strong preference for (D)Lys at the second position and (L)Asn at the third position, and binding affinity was higher than that of the arginine-rich domain of HIV-1 Tat. Thus, non-natural amino acids can yield yet additional solutions to the RNA-binding problem.

CONCLUSIONS

Recent technological advances have enhanced the capacity of peptide library screens and, together with an increasing knowledge of amino acid–RNA contacts from structural studies, have brought us closer to the design of libraries that more thoroughly encompass RNA-binding sequence space. The ability to identify RNA binders clearly depends on the nature of the RNA target site, and some binding sites are able to accommodate more diverse binding modes than others. The high frequency at which RNA-binding peptides can be found for some RNA structures suggests that specific binders may have readily evolved from a relatively small sequence space, perhaps reflecting the early transition from an RNA world. Combinatorial experiments to date indicate that polar, charged, and hydrophobic contacts all can participate in sequence-specific interactions. The results also suggest that stabilizing peptide structure, whether it be α-helix, β-sheet, or other conformations, can have a major impact on the binding affinity and specificity of short peptides. The results of further screens and new advances in peptidomimetic chemistry ultimately may be translated into novel RNA-directed therapeutics and may be expected to provide new perspectives on RNA–protein recognition.

REFERENCES

1. Patel, D. J. Curr Opin Struct Biol 1999, 9, 74–87.
2. Frankel, A. D. Curr Opin Struct Biol 2000, 10, 332–340.
3. Cheng, A. C.; Calabro, V.; Frankel, A. D. Curr Opin Struct Biol 2001, 11, 478–484.
4. Kim, S.; Ribas de Pouplana, L.; Schimmel, P. Proc Natl Acad Sci USA 1993, 90, 10046–10050.
5. Barrick, J. E.; Takahashi, T. T.; Ren, J.; Xia, T.; Roberts, R. W. Proc Natl Acad Sci USA 2001, 98, 12374–12378.
6. Barrick, J. E.; Takahashi, T. T.; Balakin, A.; Roberts, R. W. Methods 2001, 23, 287–293.
7. Barrick, J. E.; Roberts, R. W. Protein Sci 2002, 11, 2688–2696.
8. Legault, P.; Li, J.; Mogridge, J.; Kay, L. E.; Greenblatt, J. Cell 1998, 93, 289–299.
9. Austin, R. J.; Xia, T.; Ren, J.; Takahashi, T. T.; Roberts, R. W. J Am Chem Soc 2002, 124, 10966–10967.
10. Cai, Z.; Gorin, A.; Frederick, R.; Ye, X.; Hu, W.; Majumdar, A.; Kettani, A.; Patel, D. J. Nat Struct Biol 1998, 5, 203–212.

10a. Cilley, C. D.; Williamson, J. R. RNA 2003, 9, 663–676.

11. Faber, C.; Scharpf, M.; Becker, T.; Sticht, H.; Rosch, P. J Biol Chem 2001, 276, 32064–32070.
12. Scharpf, M.; Sticht, H.; Schweimer, K.; Boehm, M.; Hoffmann, S.; Rosch, P. Eur J Biochem 2000, 267, 2397–2408.
13. Tan, R.; Chen, L.; Buettner, J. A.; Hudson, D.; Frankel, A. D. Cell 1993, 73, 1031–1040.
14. Battiste, J. L.; Mao, H.; Rao, N. S.; Tan, R.; Muhandiram, D. R.; Kay, L. E.; Frankel, A. D.; Williamson, J. R. Science 1996, 273, 1547–1551.
15. Ye, X.; Gorin, A.; Ellington, A. D.; Patel, D. J. Nat Struct Biol 1996, 3, 1026–1033.
16. Tan, R.; Frankel, A. D. Proc Natl Acad Sci USA 1998, 95, 4247–4252.
17. Harada, K.; Martin, S. S.; Frankel, A. D. Nature 1996, 380, 175–179.

18. Peled-Zehavi, H.; Horiya, S.; Das, C.; Harada, K.; Frankel, A. D. RNA 2003, 9, 252–261.
19. McColl, D. J.; Honchell, C. D.; Frankel, A. D. Proc Natl Acad Sci USA 1999, 96, 9521–9526.
20. Blancafort, P.; Steinberg, S. V.; Paquin, B.; Klinck, R.; Scott, J. K.; Cedergren, R. Chem Biol 1999, 6, 585–597.
21. Weeks, K. M.; Crothers, D. M. Science 1993, 261, 1574–1577.
22. Friesen, W. J.; Darby, M. K. J Biol Chem 1997, 272, 10994–10997.
23. Friesen, W. J.; Darby, M. K. Nat Struct Biol 1998, 5, 543–546.
24. Friesen, W. J.; Darby, M. K. J Biol Chem 2001, 276, 1968–1973.
25. Harada, K.; Martin, S. S.; Tan, R.; Frankel, A. D. Proc Natl Acad Sci USA 1997, 94, 11887–11892.
26. Zhang, Q.; Harada, K.; Cho, H. S.; Frankel, A. D.; Wemmer, D. E. Chem Biol 2001, 8, 511–520.
27. Gosser, Y.; Hermann, T.; Majumdar, A.; Hu, W.; Frederick, R.; Jiang, F.; Xu, W.; Patel, D. J. Nat Struct Biol 2001, 8, 146–150.
28. Yarus, M. J Mol Evol 1998, 47, 109–117.
29. Agris, P. F.; Marchbank, M. T.; Newman, W.; Guenther, R.; Ingram, P.; Swallow, J.; Mucha, P.; Szyk, A.; Rekowski, P.; Peletskaya, E.; Deutscher, S. L. J Protein Chem 1999, 18, 425–435.
30. Mucha, P.; Szyk, A.; Rekowski, P.; Weiss, P. A.; Agris, P. F. Biochemistry 2001, 40, 14191–14199.
31. Mucha, P.; Szyk, A.; Rekowski, P.; Guenther, R.; Agris, P. F. RNA 2002, 8, 698–704.
32. Jiang, F.; Gorin, A.; Hu, W.; Majumdar, A.; Baskerville, S.; Xu, W.; Ellington, A.; Patel, D. J. Struct Fold Des 1999, 7, 1461–1472.
33. Puglisi, J. D.; Chen, L.; Blanchard, S.; Frankel, A. D. Science 1995, 270, 1200–1203.
34. Ye, X.; Kumar, R. A.; Patel, D. J. Chem Biol 1995, 2, 827–840.
35. Smith, C. A.; Calabro, V.; Frankel, A. D. Mol Cell 2000, 6, 1067–1076.
36. Ye, X.; Gorin, A.; Frederick, R.; Hu, W.; Majumdar, A.; Xu, W.; McLendon, G.; Ellington, A.; Patel, D. J. Chem Biol 1999, 6, 657–669.
37. Friedler, A.; Friedler, D.; Luedtke, N. W.; Tor, Y.; Loyter, A.; Gilon, C. J Biol Chem 2000, 275, 23783–23789.
38. Tok, J. B.; Des Jean, R. C.; Fenker, J. Bioorg Med Chem Lett 2001, 11, 43–46.
39. Tamilarasu, N.; Huq, I.; Rana, T. M. Bioorg Med Chem Lett 2000, 10, 971–974.
40. Hwang, S.; Tamilarasu, N.; Ryan, K.; Huq, I.; Richter, S.; Still, W. C.; Rana, T. M. Proc Natl Acad Sci USA 1999, 96, 12997–13002.