2562–2565 *Nucleic Acids Research, 1997, Vol. 25, No. 13*

RNA base–amino acid interaction strengths derived from structures and sequences

Brooke Lustig*, Shalini Arora and Robert L. Jernigan[1]

Department of Chemistry, San Jose State University, San Jose, CA 95192, USA and [1]Laboratory of Experimental and Computational Biology, National Cancer Institute, NIH, Building 12B, Room B116, 12 South Drive, MSC 5677, Bethesda, MD 20892-5677, USA

Received April 28, 1997 Accepted May 14, 1997

ABSTRACT

We investigate RNA base–amino acid interactions by counting their contacts in structures and their implicit contacts in various functional sequences where the structures can be assumed to be preserved. These frequencies are cast into equations to extract relative interaction energetics. Previously we used this approach in considering the major groove interactions of DNA, and here we apply it to the more diverse interactions observed in RNA. Structures considered are the three different tRNA synthetase complexes, the U1A spliceosomal protein with an RNA hairpin and the BIV TAR–Tat complex. We use binding data for the base frequencies for the seryl, aspartyl and glutaminyl tRNA–synthetase and U1 RNA–protein complexes. We compare with the previously reported DNA major groove peptide contacts the results for atoms of RNA bases, usually in the major groove. There are strong similarities between the rank orders of interacting bases in the DNA and the RNA cases. The apparent strongest RNA interaction observed is between arginine and guanine which was also one of the strongest DNA interactions. The similar data for base atomic interactions, whether base paired or not, support the importance of strong atomic interactions over local structure considerations, such as groove width and α-helicity.

INTRODUCTION

The problem of understanding RNA–protein interactions is important because RNA is more involved in function than is DNA. However, the larger structural variations manifested in RNA compared to DNA make its study more difficult. The complexity of the problem resembles the case of protein–protein interactions which has met with some recent success using approaches (1–3) similar to that taken below. In some ways, the difference between DNA and RNA binding sites corresponds to a difference in dimensionality—DNA double helical structure is nearly one dimensional. On the other hand, RNA presents highly variable surfaces for interaction, more similar to protein structures. One way of comprehending the complexities of such structures is to deconstruct them in terms of the interactions that stabilize them. Furthermore, if there were sufficiently large numbers of diverse structures, then the effects of both the RNA and the protein structure would be averaged out, and the dominant atomic interactions would become evident. Here we are going to compile and analyze the structural data available for RNA–protein structures to learn about their interactions. This will be done at a coarse grained level of base–amino acid pairs rather than detailed individual atomic pairs. The principal difficulty in learning about RNA–protein interactions remains the fact that there are relatively few available structures.

Interactions between RNA and proteins ought to provide a variety of interactions similar to that between pairs of proteins. For proteins there are potentially 20 types of residues to interact compared to the four nucleotide bases. Two questions arise. How do the RNA–protein interfaces achieve a comparable level of diversity for recognition? Are there important structural motifs for binding between proteins and RNA? The search for protein binding motifs has proven to be almost pointless, since a wide variety of protein structure elements are now known to interact with DNA. The much greater diversity of RNA structures would seem to make the dominance of only a few structural motifs even less likely. However, this does not preclude the occurrence of dominant motifs of atomic interactions. So, for RNA–protein binding, we are simply going to look at the frequencies of base–amino acid interactions, without consideration of their structural context.

How do the structural differences between RNA and DNA affect protein binding? In RNA there are additional features in comparison to DNA, that facilitate specific recognition by proteins. The additional G·U base pair type, beyond the canonical A·U and G·C types of base pairs, adds to the diversity of possible interactions. Furthermore, the available RNA structures already show a remarkable variety of other ways in which bases can hydrogen bond to one another, e.g., triplets, purine–purine and pyrimidine–pyrimidine pairs. In addition, there are unpaired bases in bulges and loops that can interact with amino acids. So, the bases themselves do offer a rich diversity for binding to the 20 types of amino acids. Also some amino acids are capable of interacting simultaneously with several base pairs so this provides a further variety on the RNA surface for protein interactions. Overall, this catalog of potential interacting RNA surface features affords a sufficient number of ways to achieve their specific recognition by proteins. But, as we will see below highly favorable interactions can dominate amino acid–base interactions.

* To whom correspondence should be addressed. Tel: +1 408 924 4968; Fax: +1 408 924 4945; Email: lustig@batnet.com

For a given protein binding site on RNA, how variable can its sequences be? From analyses of DNA–protein binding sequences (4), it appears that the strengths of individual interacting pairs are not so critical. DNA binding sequence frequencies indicate that some interacting bases can be replaced. In part, this may reflect the replacement of one hydrogen bond acceptor or donor by a similar one from another base. However, there is also the possibility that substitutions can be energetically compensating, i.e., a more weakly binding base might be acceptable, if another simultaneous base substitution elsewhere in the binding site were made with a stronger binder. Are RNA–protein binding sequences similarly variable?

The advent of sequence libraries for selecting active binding sequences is having a major impact on the study of these systems. The present approach could be applied directly to assist in the design of better combinatorial libraries. Other approaches applying pattern recognition methods are being developed to design and analyze combinatorial libraries for peptides and related polymers (5,6). Another useful approach has been to examine and analyze nature's functional combinatorial libraries by aligning and determining DNA base preferences from variant sequence data (7).

Others have also been cataloging the interactions found from the limited set of three-dimensional DNA–protein structures as determined by X-ray crystallography (8, Mandel-Gutfriend,Y., Margalit,H., Jernigan,R.L. and Zhurkin,V.B., personal communication). But, the present approach goes beyond the strictly structural to include additional information both from binding data and from sequence variability. We have previously derived self-consistent normalized relative energies for each of the four DNA bases interacting in the major groove with a specific amino acid (4) by using an extensive set of data collected from combinatorial multiplex DNA binding of zinc finger domains (9). The five strongest interactions found were: Lysine · guanine, Lysine · thymine, Arginine · guanine, Aspartic acid · cytosine and Asparagine · adenine. These relative energies correlated well with those derived from DNA binding data for Cro and λ repressors and the R2R3 c-Myb protein domain (10–12), as well as similar interaction energies derived directly from frequencies of bases determined to be in contact with particular amino acids in the bacteriophage λ operator sequences.

A major objective of the present work is to calculate RNA–protein potentials. Those for major groove interactions can be compared directly with those derived for the major groove of DNA. RNA differs from DNA in some ways, but since we consider multiple structures as well as only relative values among the four bases, many differences such as those arising from the greater stiffness of the RNA backbone relative to DNA might be important. There are still some remaining differences but the present considerations are only semi-quantitative, and we will be comparing only the strongest effective interactions. The present considerations include data from BIV TAR–Tat binding where NMR was used to identify specific base–amino acid contacts (13,14). In addition we use the more extensive data for RNA base sequence frequencies of acyl tRNAs and U1 RNAs (15,16) that are identified by X-ray at positions in specific contact with particular amino acids (17–22).

METHODS AND RESULTS

We use frequencies of contacts between bases and amino acids to derive relative interaction energies from the acyl tRNA–synthetase and U1A spliceosomal protein–RNA complexes, as we did earlier for zinc fingers interacting with DNA (4). First we calculate the logarithms of frequencies for all occurrences of a j-type base interacting specifically with an I-type amino acid so that the interaction energy e_{Ij} is of the form

$$e_{Ij} \sim -\ln f_{Ij} \qquad \mathbf{1}$$

where f_{Ij} is the sum over all the sets of the relative frequencies in which a base type j interacts with all occurrences of a residue type I. For each of the four bases, the relative interaction energies are then normalized as

$$\sum_j \ln f_{Ij} = 0 \qquad \mathbf{2}$$

This corresponds to a reference state that shifts the values so that the mean for the four bases is zero.

The U1A spliceosomal protein–RNA hairpin structure from X-ray indicates 14 base–amino acid contacts at Arginine52 · G16, Arginine52 · A6, Glutamic acid19 · U7, Asparagine16 · G9, Asparagine15 · G9, Lysine80 · U8, Asparagine16 · U8, Glutamine85 · C10, Tyrosine86 · C10, Lysine88 · C10, Aspartic acid92 · C12, Serine91 · A11, Threonine89 · A11, and Aspartic acid90 · C12 (22); these include both peptide side chain and peptide backbone interactions. The corresponding collected frequencies for the RNA sequences (16) are utilized. The binding domain of the protein is primarily at the loop A6 through C15 of the 21 base synthetic RNA hairpin loop. The amino acid types of the contacts identified by X-ray are considered here to be conserved (23). The sets of four relative base–amino acid energies are explicitly derived using equations **1** and **2**. Stacking and hydrophobic interactions have not been considered here. The only major groove contact was reported for Arginine52 · G16.

We have used similar sequence data (15,24) and structures for the seryl (17), aspartyl (18) and glutaminyl (19,20) tRNA–synthetase complexes. They present a more diverse set of interactions than the U1A spliceosomal protein with RNA, since almost half are anticodon loop contacts involving Glutamic acid188 · U^{Asp}34, Arginine119 · U^{Asp}35, Glutamine138 · U^{Asp}35; Alanine414 · C^{Gln}34, Arginine341 · U^{Gln}35, Glutamine517 · U^{Gln}35, Arginine520 · U^{Gln}35 and Arginine402 · G^{Gln}36. The remaining contacts are found at Glutamic acid327 · G^{Asp}73, Asparagine330 · A^{Asp}72; Alanine555 · G^{Ser}19, Glutamine545 · G^{Ser}47a, Glutamine545 · G^{Ser}47n and also include the two major groove contacts at Asparagine330 · U^{Asp}1 and Asparagine330 · A^{Asp}72. This calculation also includes in the same way the data for the U1 RNA–protein case.

Relative interactions for individual amino acids with the four bases are shown in Figure 1 for the combined data from the U1 RNA–spliceosomal protein and tRNA–synthetase structures and sequences. Also shown for comparison are the strong DNA interactions derived previously for major groove interactions. The RNA cases include diverse interactions, and only the first two cases designated by Rm and Nm are for major groove interactions. For RNA major groove interactions, the most favored pair is arginine with guanine which was also one of the five strongest pairs for DNA. In the case of non-major groove interactions in RNA, other strongest cases for interaction that can be seen in this figure are Arginine · uridine, Lysine · cytosine, Aspartic acid · cytosine, Glutamic acid · guanine, Alanine · guanine, Tyrosine · cytosine and Serine · or Threonine · adenosine.

We can reasonably assume that the aspartyl, seryl and glutaminyl synthetase amino acid contacts are conserved, and we show

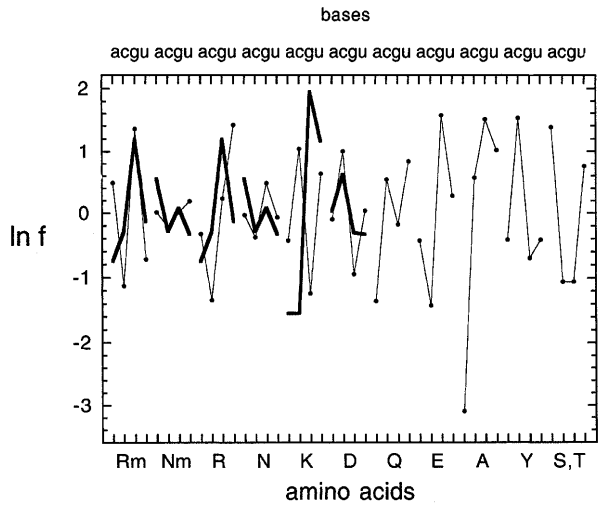

Figure 1. Logarithms of the relative frequencies, i.e. relative energies, of base–amino acid interactions. The logarithms are normalized over the four bases. The amino acid type is specified on the lower abscissa, and the base type is given along the top abscissa. In the DNA case base 'u' should be interpreted as t. DNA data from our previous work is shown as bold solid lines without points; the present RNA data is given by points connected by thin lines. RNA data is calculated from the base and amino acid frequencies in 39 U1 RNAs (16) at base–amino acid contacts identified from the X-ray crystal structure for the U1A spliceosomal protein complex (22). Also included are data calculated for sequences (15) at base–amino acid contacts identified from X-ray crystal structures for glutaminyl (18 sequences), aspartyl (20 sequences) and seryl (42 sequences) synthetase–tRNA complexes (17,18,20,21,24).

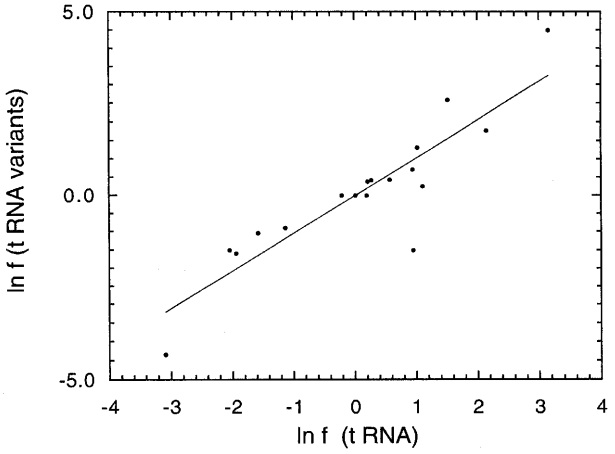

Figure 2. Species variability of RNA base–amino acid interactions in tRNA–synthetase complexes. Correlations are shown for relative base–amino acid interaction energies averaged over all RNA sequences of glutaminyl, aspartyl and seryl tRNA–synthetase complexes (abscissa) with those of variants within individual species (ordinate). The ordinate values have been averaged first over all occurrences for one species and subsequently averaged over all species. Twenty points are included for the four amino acids arginine, glutamine, glutamic acid and alanine in any position other than major groove, and asparagine in the major groove. The straight line fit has an intercept of zero and a slope of 1.03; the correlation coefficient is 0.89.

(Fig. 2) that there is a correlation between interaction energies derived for RNA sequence variants in individual species where the protein is constant compared to values derived directly from sequences for all species. The correlation coefficient of 0.89 for 20 points indicates a probability of being random as <0.001 (30).

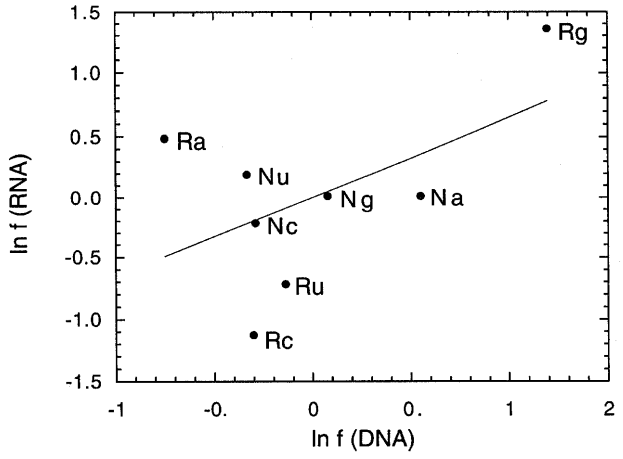

Figure 3. Comparison of the strongest interactions in DNA (4) and in RNA, for interactions with major groove atoms. RNA results are the relative energies for U1 RNA–protein (arginine) together with tRNA–protein (asparagine). This includes all data from Figure 1 given as Rm and Nm together with the R and N data not in major grooves but interacting with base atoms that are usually accessible in major grooves. Relative interaction energy points are labeled with upper case amino acid and lower case base. The straight line fit has zero intercept and a slope of 0.93; the correlation coefficient is 0.5; the probability of being random is <0.21.

Both asparagine contacts that are not in the major groove occur in glutaminyl tRNA, and involve unusual base–amino acid contacts with non-Watson–Crick-modified base pairs (20). These have not been included in the present analyses. It is clear in Figure 1 that the major groove interactions for arginine and asparagine differ from the corresponding non-major groove interactions. Most interestingly for these two amino acid types, the corresponding relative interaction energies determined for major groove RNA base–amino acid interactions are quite strongly correlated with the interactions derived from DNA base–amino acid major groove contacts (Fig. 3) where we have included not only the cases in the major groove category but also cases where the interactions are with the atoms usually accessible in the major groove. The data is fit with a straight line of slope of 0.93 and a zero intercept. The correlation coefficient of 0.5 for the eight points indicates that the probability of being random is <0.2 (30). Clearly it can be seen that the results are better for asparagine than for arginine. The non-major groove RNA base–amino acid interactions show no significant correlation with results for the strongly binding amino acids of major groove DNA, with the exception of aspartic acid and asparagine (Fig. 1). When only the more restricted major groove cases were considered using only the Rm and Nm data in Figure 1, similar results are observed with a slope of 0.66 and a correlation coefficient 0.53. The total number of data points for the less restricted case for asparagine is doubled and the number for arginine is quadrupled. The similarity in the results lends strong support to focusing on atomic interactions rather than nucleic acid structure.

Also there are three polar RNA base–amino acid contacts identified by NMR for BIV TAR–Tat, Arginine70·G14, Arginine73·G11 and Glycine71·G22 (13). All of these contacts, for arginine and glycine, are contacts in the major groove. We calculate normalized relative interaction energies from binding studies of BIV Tat peptide (13) and various mutants of BIV–TAR assuming that f_{Ij} is proportional to the corresponding binding

constant. That data is incomplete, since not all four bases have been substituted. There is still some useful ranking information about base–amino acid binding. For arginine there is a clear preference for guanine over cytosine, which is consistent with the previous RNA and DNA results. And for glycine, the interaction with adenine is stronger than for cytosine.

DISCUSSION

It is noteworthy that there is a clear correlation between the relative interaction energies for the DNA and RNA major groove atomic contacts with arginine and asparagine. This suggests, given their importance in DNA base–amino acid interactions (4), that simple charge or hydrogen bond considerations are the explanation for the sequence dependence of RNA base–amino acid contact preferences rather than a dependence on the RNA structure. Inspection of the interaction energies is generally consistent with simple base and amino acid charge considerations. Also it is significant that the relative interaction energies for exclusively non-major groove RNA base–amino acid contacts appear to be completely different in character from those associated with major groove DNA and RNA. This is consistent with the lack of specificity noted for minor groove interactions in DNA (31) and RNA (19).

Focusing on the more extensive data in Figure 1 for RNA, the dominant major groove interaction is Arginine·guanine and for the non-major groove cases Arginine·uracil, Lysine·cytosine, Aspartic acid·cytosine, Glutamic acid·guanine, Alanine·guanine, Tyrosine·cytosine and Serine/Threonine·adenine. The importance of arginine in specific and non-specific binding has been previously noted (32). We limit the category of specific interactions to base–amino acid contacts. We have not considered other non-specific interactions involving phosphates and riboses here because preliminary analyses showed no significant DNA or RNA sequence specificity (Lustig,B., unpublished results; 33). Shi and Berg (33) have shown that zinc finger RNA does not differ from zinc finger DNA in sequence, but in RNA has an enhanced binding which may involve increased interactions with phosphates or 2′-OHs. Our results suggest that there are diverse ways to obtain specific interactions but that there are some dominant interactions. Remarkably, several of these stand out already in the present limited data. It must be noted, however, that the interactions present in the structures here are likely to provide an incomplete list of all RNA interactions.

The specific occurrence of Arginine·guanine pairs in several recent structures (25–29) is noteworthy. The occurrences of this pair in diverse structural contexts is particularly important. For example, in the HIV-2 Tar–Argininamide complex (25), this pair occurs with the argininamide stacked between U and A bases where the U is also involved in a U·A·U triplet. In other cases arginine was shown even to cause conformational transitions in DNA (26). And, in another RNA study (29), arginines which were originally in α-helices still bind even when the helices have been disrupted by changing the peptide sequences. This suggests that the arginine pair of hydrogen bonds formed with N7 and O6 of guanine is extremely strong. Perhaps, these are sufficiently strong that they form in spite of the structural context. The present approach can readily treat this class of strong interactions, even when the data are limited. In other less favored cases, averaging over sufficient numbers of structures is required.

A clearer, more precise elucidation of RNA base–amino acid interactions requires a more extensive set of structures or experimental data such as those that could be derived from combinatorial RNA–protein binding studies for a variety of well characterized three-dimensional structures. Ultimately the present type of results could be utilized for sequence design in a variety of problems. If a given surface region of protein were targeted for binding to a new RNA, then the protein sequence could be utilized directly to suggest the composition of RNAs that would be likely to bind most specifically. Sequences with such a composition could be screened experimentally with an appropriately designed RNA combinatorial library (34).

REFERENCES

1 Wallqvist,A., Jernigan,R.L. and Covell,D.G. (1995) *Protein Sci.* **4**, 1881–1903.
2 Laskowski,R.A., Thornton,J.M., Humblet,C. and Singh,J. (1996) *J. Mol. Biol.* **259**, 175–201.
3 Miyazawa,S. and Jernigan,R.L. (1996) *J. Mol. Biol.* **256**, 623–644.
4 Lustig,B. and Jernigan,R.L. (1995) *Nucleic Acids Res.* **23**, 4707–4711.
5 Siani,M.A., Weininger,D. and Blaney,J.M. (1994) *J. Chem. Inf. Comput. Sci.* **34**, 588–593.
6 Martin,E.J., Blaney,J.M., Siani,M.A., Spellmeyer,D.C., Wong,A.K. and Moos,W.H. (1995) *J. Med. Chem.* **38**, 1431–1436.
7 Schneider,T.D., Stormo,G.D., Gold,L. and Ehrenfeucht,A. (1986) *J. Mol. Biol.* **188**, 415–431.
8 Mandel-Gutfreund,Y., Schueler,O. and Margalit,H. (1995) *J. Mol. Biol.* **253**, 370–382.
9 Desjarlais,J.R. and Berg,J.M. (1994) *Proc. Natl. Acad. Sci USA* **91**, 11094–11103.
10 Takeda,Y., Sarai,A. and Rivera,V.M. (1989) *Proc. Natl. Acad. Sci. USA* **86**, 439–443.
11 Sarai,A. and Takeda,Y. (1989) *Proc. Natl. Acad. Sci. USA* **86**, 6513–6517.
12 Ogata,K., Kanei-Ishii,C., Sasaki,M., Hatanaka,H., Nagadoi,A., Enari,M., Nakamura,H., Nishimura,Y., Ishii,S. and Sarai,A. (1996) *Nature Struct. Biol.* **3**, 178–187.
13 Chen,L. and Frankel,A.D. (1994) *Biochemistry* **33**, 2708–2715.
14 Puglisi,J.D., Chen,L., Blanchard,S. and Frankel,A.D. (1995) *Science* **270**, 1200–1203.
15 Sprinzl,M., Steegborn,C., Hubel,F. and Steinberg,S. (1996) *Nucleic Acids Res.* **24**, 68–72.
16 Gu,J. and Reddy,R. (1996) *Nucleic Acids Res.* **24**, 73–75.
17 Biou,V., Yaremchuk,A., Tukalo,M. and Cusack,S. (1994) *Science* **263**, 1404–1410.
18 Cavarelli,J., Rees,B., Ruff,M., Thierry,J.-C. and Moras,D. (1993) *Nature* **362**, 181–184.
19 Ruff,M., Krshnaswamy,S., Boeglin,M., Poterszman,A., Mitschler,A., Podjarny,A., Rees,B., Thierry,J.C. and Moras,D. (1991) *Science* **252**, 1682–1689.
20 Rould,M.A., Perona,J.J., Soll,D. and Steitz,T.A. (1989) *Science* **246**, 1135–1142.
21 Rould,M.A., Perona,J.J. and Steitz,T.A. (1991) *Nature* **352**, 213–218.
22 Oubridge,C., Ito,N., Evans,P.R., Teo,C.-H. and Nagai,K. (1994) *Nature* **372**, 432–438.
23 Watson,J.D., Hopkins,N.H., Roberts,J.W., Steitz,J.A. and Weiner,A.M. (1987) *Molecular Biology of the Gene, Fourth Ed.* Benjamin-Cummings, Menlo Park, CA.
24 Singhal,R.P. and Fallis,P.A.M. (1979) *Progr. Nucleic Acid Res. Mol. Biol.* **23**, 227–290.
25 Brodsky,A.S. and Williamson,J.R. (1997) *J. Mol. Biol.* **267**, 624–639.
26 Harada,K. and Frankel,A.D. (1995) *EMBO J.* **14**, 5798–5811.
27 Tao,J. and Frankel,A.D. (1992) *Proc. Natl. Acad. Sci. USA* **89**, 2723–2726.
28 Battiste,J.L., Mao,H., Rao,N.S., Tan,R., Muhandiram,D.R., Kay,L.E., Frankel,A.D. and Williamson,J.R. (1996) *Science* **273**, 1547–1551.
29 Tau,R. and Frankel,A.D. (1995) *Proc. Natl. Acad. Sci. USA* **92**, 5282–5286.
30 Bevington,P.R. (1969) *Data Reduction and Error Analysis for the Physical Sciences*. McGraw-Hill Book Co., NY, p. 311.
31 Seeman,N.C., Rosenberg,J.M. and Rich,A. (1976) *Proc. Natl. Acad. Sci. USA* **73**, 804–808.
32 Burd,C.G. and Dreyfus,G. (1994) *Science* **265**, 615–621.
33 Shi,Y. and Berg,J.M. (1995) *Science* **268**, 282–284.
34 Tuerk,C. and Gold,L. (1990) *Science* **249**, 505–510.