Proc. Natl. Acad. Sci. USA
Vol. 86, pp. 7402–7406, October 1989
Biochemistry

RNA structure, not sequence, determines the 5′ splice-site specificity of a group I intron

(ribozyme/self-splicing intron/exon–intron junction)

JENNIFER A. DOUDNA, BRENDAN P. CORMACK, AND JACK W. SZOSTAK

Department of Molecular Biology, Massachusetts General Hospital, Boston, MA 02114

Communicated by Tom Maniatis, July 10, 1989

ABSTRACT **The group I self-splicing introns act at exon–intron junctions without recognizing a particular sequence. In order to understand splice-site selection, we have developed an assay system based on the *Tetrahymena* ribozyme to allow the study of numerous 5′-splice-site variants. Cleavage at the correct site requires formation of the correct secondary structure and occurs most efficiently within a 3-base-pair window centered on base pair 5 from the bottom of the P1 stem. Within this window the ribozyme recognizes and cleaves at a "wobble" base pair; the base pair above the cleavage site also influences splicing efficiency. The recognition of RNA structure rather than sequence explains the ability of these transposable introns to splice out of a variety of sequence contexts.**

The self-splicing rRNA intron of *Tetrahymena* catalyzes three site-specific transesterification reactions that result in ligation of the exons, excision, and circularization of the intron (1–3). Although a great deal has been learned about the structure and activity of this ribozyme in recent years, the mechanism of splice-site recognition by the intron has remained elusive. Because there are no constant sequences at the exon–intron junctions, it is clear that the ribozyme must recognize some aspect of RNA structure rather than nucleotide sequence.

The 5′ splice site has been determined for many group I introns, all of which share sequence homology, secondary structure, and presumably a common catalytic mechanism with the *Tetrahymena* intron. Early comparisons of the secondary structures of several group I introns revealed that the 5′ splice site always coincided with the position of a U·G base pair in the first stem–loop (P1) of the intron (4–7). This has held true for all group I introns subsequently identified. Mutational analysis has shown that the integrity of this stem is essential for catalysis: single mutations in the stem greatly reduce splicing activity, but compensatory base changes that restore base pairing also restore activity (8, 9). Mutations that change either base of the U·G base pair have also been found to greatly decrease splicing (9).

We have used an assay system derived from the *Tetrahymena* intron to investigate the structural features of the 5′ exon–intron junction region that are necessary for cleavage by the intron core (Fig. 1). The ribozyme we used is a shortened version of the intron core containing all of the conserved sequences and secondary structures (11). When incubated with a substrate RNA molecule containing the 5′ exon–intron junction and flanking sequences, this ribozyme cleaves the substrate at the expected splice site (11, 12). The association of intron core and substrate partially regenerates a base-paired stem and thus reconstitutes the structure of the intact intron from two fragments (J.A.D. and J.W.S., unpublished data). An advantage of this system is that the substrate

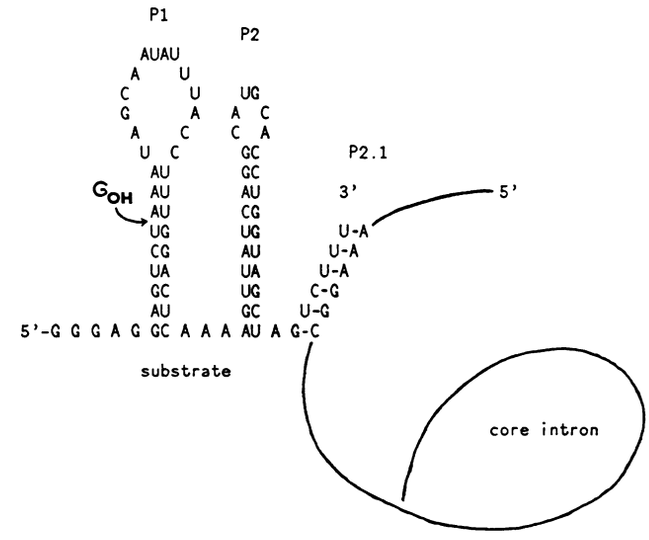

FIG. 1. Trans-splicing reaction. The reaction is analogous to the first step of self-splicing. The substrate consists of the two hairpins P1 and P2 and part of P2.1. The intron core extends from the 3′ side of the P2.1 stem through P7 (see ref. 10 for nomenclature definitions). When substrate and intron are incubated together in the presence of Mg^{2+} and free guanosine, the intron catalyzes guanosine attack at the indicated site in P1 of the substrate. The guanosine becomes covalently attached to the 5′ end of the RNA at that position; the short 5′ fragment of the substrate is the leaving group.

is small enough to be made by T7 RNA polymerase transcription of synthetic oligonucleotides, so that variants can be rapidly designed and tested. We have used such substrate variants to define those aspects of the structure of the P1 stem that contribute to 5′ splice-site specificity.

MATERIALS AND METHODS

Intron Mutants. Mutants were constructed by cassette mutagenesis of plasmid pAG100, a pBR322 derivative containing the modified *Tetrahymena* intron downstream of the bacteriophage T7 RNA polymerase promoter (11). Vector was prepared by digesting pAG100 with restriction endonucleases *Eco*RI and *Sal* I. The small *Eco*RI–*Sal* I fragment was replaced by synthetic 68-mer oligonucleotides that contained the desired point mutation. Plasmid DNA was prepared and sequenced through the insert by the method of Sanger *et al.* (13). Clones with the correct sequence were digested with *Bam*HI and transcribed *in vitro* with T7 RNA polymerase. RNA was purified by electrophoresis in 6% polyacrylamide gels containing 7 M urea. After elution, phenol extraction, and ethanol precipitation, RNA was resuspended in deionized water at a concentration of 1 μM.

Substrate Mutants. Oligonucleotides were synthesized on a Biosearch model 8750 DNA synthesizer and purified by

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

urea/polyacrylamide gel electrophoresis. RNA was synthesized by runoff transcription of oligonucleotides (11, 14). RNA was purified on 40-cm 20% polyacrylamide gels containing 7 M urea. This allowed separation of the correct-sized RNA from the $n \pm 1$ products that arise during transcription. After elution, phenol extraction, and ethanol precipitation, RNA was resuspended in deionized water at a concentration of 10 μM.

Activity Assays. Trans-splicing assays were performed using 0.4 μM intron, 1 μM substrate, and 2 μM [α-^{32}P]GTP in a buffer containing 30 mM Tris/HCl (pH 7.4), 20 mM $MgCl_2$, 10 mM NH_4Cl, and 0.2 mM aurin tricarboxylic acid. Reaction mixtures (10 μl) were incubated at 45°C for 20 min, then mixed with 10 μl of loading dye containing 90% (vol/vol) formamide, 10 mM Tris/HCl (pH 7.5), 1 mM EDTA, 0.4% xylene cyanol, and 0.4% bromophenol blue. Reaction products were analyzed by denaturing polyacrylamide gel electrophoresis and autoradiography. Quantitation of band intensity was done directly with a Betagen blot analyzer (Betagen, Waltham, MA), prototype model. Reaction conditions for K_m determinations were as above except that the ribozyme concentration was 0.05 μM, GTP concentration was 100 μM, and substrate concentration was varied from 0.1 μM to 5.0 μM. These reaction conditions are within the linear range of a time course and correspond to much less than one turnover per enzyme molecule.

RESULTS

The substrate originally used in the trans-splicing experiments was a 119-nucleotide RNA consisting of 25 nucleotides of vector sequence at the 5′ end followed by 32 nucleotides of the 5′ exon and 62 nucleotides of the intron (12). By stepwise deletion, the substrate was reduced to a 73-nucleotide species that retained activity as a substrate (11). This substrate was spliced at several sites due to the occurrence of alternative secondary structures. Formation of these alternative secondary structures could be prevented, and correct splice-site specificity restored, by changing the three base pairs at the bottom of the P1 stem (11). We have used this substrate (Fig. 1) as the parent molecule for the base changes and deletions described in this paper.

Position of the U·G Base Pair Within the P1 Stem Is Important. In all known group I introns, the 5′ exon–intron junction falls within a stem–loop referred to as P1. Neither the sequence nor the length of P1 is phylogenetically conserved. The only obviously conserved feature of P1 is that the base at the 3′ end of the exon is invariably a uracil in the 5′ strand of the P1 stem that is paired with a guanine residue in the 3′ strand. In approximately 50% of the group I introns we examined, the U·G base pair is the sixth base pair from the bottom of the P1 stem. This weak conservation suggested that the position of the U·G might be significant.

To find out how far up and down the U·G could be moved within the P1 stem and still allow cleavage to occur after the U, we synthesized substrates with a U·G base pair at position 1, 2, or 3 above or 1, 2, or 3 below the position of the original U·G base pair (Fig. 2*A*). In these substrates, the naturally occurring U·G was replaced by an A·U base pair (changing the U·G at its normal position to A·U results in a substrate that is cleaved very poorly by the ribozyme). When the U·G is one base pair below its normal position (i.e., 5 base pairs from the bottom of P1), cleavage is almost as efficient as with wild-type substrate. Moving the U·G 1 position up or 2 positions down from its usual location leads to substantially less efficient cleavage, and moving the U·G 2 or 3 positions above or 3 positions below its usual location abolishes activity. Thus, the ribozyme recognizes the U·G base pair, but only when it is located within a narrow positional window in the P1 stem. This is consistent with the fact that cleavage

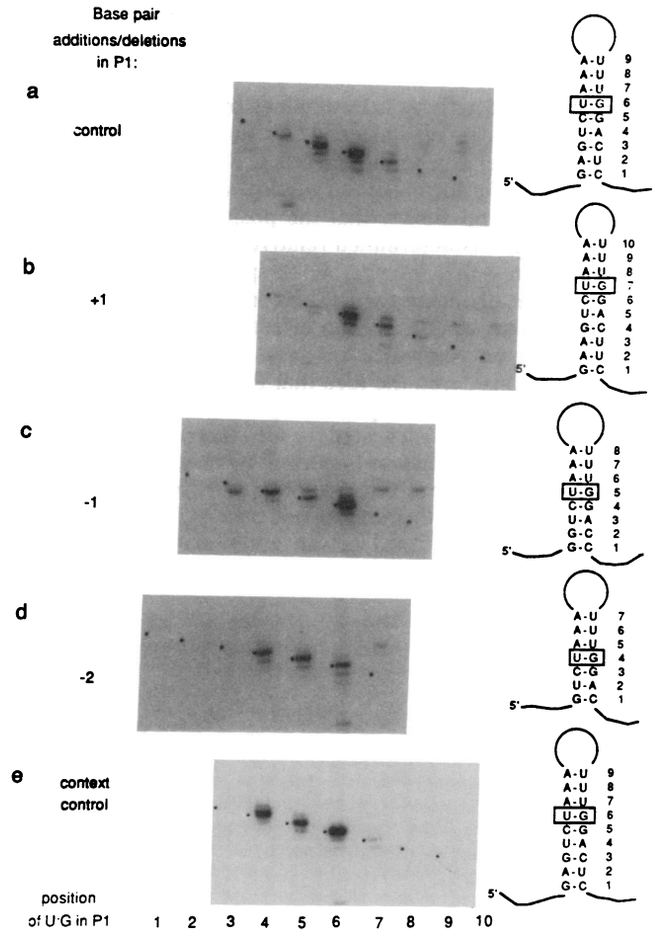

FIG. 2. The window of splicing activity in P1. For each panel a set of seven substrates was synthesized as described in *Materials and Methods*. The P1 stem of the parent molecule for each set is shown at right. Each set was tested for splicing by the intron core. Dots highlight the band representing cleavage at the U·G. The other bands represent low-level cutting at other base pairs. The major secondary band represents incorrect splicing at the C·G base pair found just below the normal U·G position. (*a*) Derivatives of the substrate shown in Fig. 1 were made in which A·U replaces U·G and the U·G is moved up or down 1, 2, or 3 positions, generating a family of substrates in which the U·G is 3–9 base pairs from the bottom of the stem. Efficient cleavage is seen when the U·G is at position 5 or 6, and weak cleavage when it is at position 4 or 7. (*b*) An A·U base pair was inserted at position 2 of P1; thus, in the parent molecule for this set the U·G is 7 base pairs from the bottom of the stem. A family of substrates in which the U·G is 4–10 base pairs from the bottom of the stem was generated as above. Efficient cleavage is seen when the U·G is at position 6; weak cleavage is seen when it is at position 4, 5, 7, or 8. (*c*) In this family of substrates, base pair 2 of P1 was deleted so that the parent molecule of this set has the U·G base pair 5 positions from the bottom of the stem. Moving the U·G generates a family of substrates in which the U·G is 2–8 base pairs from the bottom of the stem. Cleavage at the U·G is seen only when the U·G is in position 4, 5, or 6. (*d*) In this family, base pairs 2 and 3 of P1 were deleted, so that in the parent molecule the U·G is 4 base pairs from the bottom. Moving the U·G generates a family in which the U·G is 1–7 base pairs from the bottom of the stem. Correct splicing is seen only when the U·G is in position 4, 5, or 6. The autoradiogram is overexposed; activities of this family of substrates are about 10-fold lower than those of the other families, presumably because of the low stability of the short stem. An analogous family of substrates in which the stem is stabilized by the addition of an extra A·U base pair to the top of the stem shows the same pattern but higher levels of splicing (data not shown). (*e*) In this family, with the U·G at positions 2–9, only A·U base pairs are present above the U·G, to minimize sequence context effects. Maximal activity is seen for positions 4, 5, and 6.

is never observed at a U·G base pair found naturally in P1 at a position 4 base pairs below the U·G that marks the exon–intron junction.

P1 Window of Activity Is Determined by Counting from the Bottom of the Stem. We imagined that some constraint on the position of P1 relative to the active site of the ribozyme must be responsible for the observed window of activity. First, we considered the possibility that the position of the cleavage window in P1 was determined relative to the position of the P2 stem. To test this, we altered the spacing between P1 and P2. However, changing the P1/P2 connecting segment from AAA to AAAA or AA had no effect on the P1 window of activity (data not shown).

We then considered the idea that P1 itself interacted with the ribozyme in such a way that only certain base pairs could reach the active site. We tested the idea that the distance from the bottom of the stem was important by making three sets of substrates. In the first set, one base pair (base pair 2) was deleted from P1; in the second set, base pairs 2 and 3 were deleted; and in the third set an additional base pair was inserted above base pair 1. In each set of substrates, the U·G was moved to seven positions in the stem. In all three families of substrates, optimal activity was found when the U·G base pair was the fourth, fifth, or sixth base pair from the bottom of the stem (Fig. 2 *b–d*). For example, adding one base pair to the bottom of P1 moves the U·G from position 6 to position 7, where it is cleaved poorly; activity is restored in this set by moving the U·G back down the stem to position 6. Cleavage is not observed at position 3 or lower in any of the sets, but cleavage is observed in all sets at positions 4 and 5. The variable extent of cleavage at positions 4 and 5 is in large part due to sequence context effects (see below). To minimize context effects and to show the window effect more clearly, activities of a set of substrates with the U·G in different positions and only A·U base pairs above it in the stem were compared (Fig. 2*e*).

To test the possibility that distance from the top of the stem might also be important, two additional families of substrates were synthesized. These substrates, with one base pair added or deleted from the top of P1, showed no effect on the window of activity (data not shown). These experiments clearly show that the activity of a given substrate is primarily determined

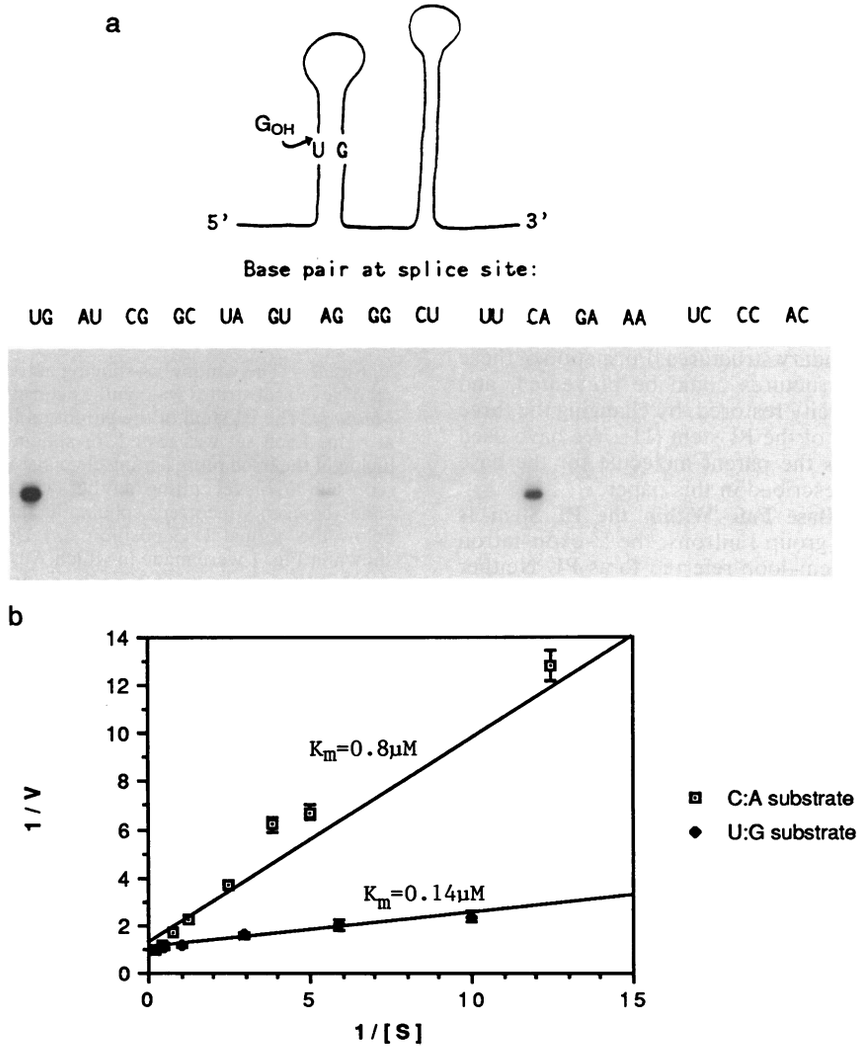

FIG. 3. Activities of substrates with base-pair changes at the splice site in P1. (*a*) Substrates with the 16 different base combinations at the splice site in P1 were constructed. The parent molecule for these mutants is shown in Fig. 1. The substrates were tested for splicing by the intron under our standard splicing conditions and analyzed by urea/polyacrylamide gel electrophoresis. An autoradiogram of a sample gel is shown. The major band corresponds to the 63-nucleotide expected spliced product. (*b*) K_m values were determined for the substrates with either a U·G or a C·A base pair at the splice site. Lineweaver–Burk plots are shown; each point is the average of three determinations. K_m(U·G) = 0.14 ± 0.02 μM; K_m(C·A) = 0.8 ± 0.1 μM. Kinetic parameters for the other substrates could not be obtained because of the low levels of products and because multiple products are formed due to cleavage at multiple sites within the recognition window in P1.

by the position of the U·G relative to the bottom of the P1 stem.

C·A Can Partially Substitute for U·G. Although the window of activity contributes to 5′ splice-site specificity, recognition of the U·G base pair within the window augments this specificity. How is the U·G base pair recognized? The phylogenetic data are not helpful in this case because the U·G is a completely invariant feature of the group I introns. We synthesized a set of substrates in which the splice-site U·G was replaced by every possible base combination at the splice site. The products of reactions with this complete set of substrates are shown in Fig. 3*a*. Remarkably, cleavage of the C·A combination is almost 50% of that of the wild-type U·G. The other 14 substrates show only low-level cleavage by the ribozyme, some at slightly variant positions. We examined the effect of the U·G → C·A change on substrate binding by measuring the K_m values of these substrates with wild-type ribozyme. Fig. 3*b* shows Lineweaver–Burk plots for the two substrates. The K_m for the wild-type substrate is 0.14 ± 0.02 μM, while the K_m for the C·A substrate is 0.8 ± 0.1 μM. Since the ribozyme is 50% less active with the C·A substrate than with the wild-type substrate even at saturating substrate concentrations, the U·G → C·A change results in a decrease in k_{cat} as well as an increase in K_m. The slightly weaker binding and activity of the C·A substrate (K_m of 0.8 μM, vs. 0.14 μM for the U·G substrate) may reflect the lower stability of the C·A base pair (Fig. 4).

The Base Pair Above the U·G Is Important. The base pairs flanking the U·G pair at the splice site are highly variable in the group I introns. Changes in P1 that retain base complementarity can have small but definite effects on splicing (8, 9, 15). To systematically examine the roles of the base pairs flanking the U·G, we synthesized substrates with the base pair either 1 position above or 1 or 2 positions below the U·G changed to the other three Watson–Crick base pairs. The results from these experiments are summarized in Table 1. Most base-pair changes have little effect on splicing efficiency, at either 45° or 58°C. The strongest effects are seen with the base pair above the U·G. When the U·G is at position 6, a C·G at position 7 results in an 80% decrease in activity. This effect is also seen when the U·G is at other positions, and accounts for the low level of splicing seen when the U·G is moved to position 4 in our standard P1 stem (Fig. 2*a*).

Table 1. Activities of substrates with base-pair changes in P1 above and below the U·G splice site

Position in stem	Relative activity			
	G·C	C·G	A·U	U·A
7	0.6	0.2	*1.0*	0.5
5	0.9	*1.0*	1.0	1.2
4	0.8	0.6	1.0	*1.0*

Derivatives of the substrate shown in Fig. 1 were constructed and tested for splicing. The italicized numbers in the table correspond to the normal base pair at that position. Activities are the average of three determinations.

DISCUSSION

Our results show that there are at least four aspects of RNA structure that contribute to splice-site specificity. (*i*) The RNA must fold into the proper secondary structure so that the P1 stem forms; alternative secondary structures result in cleavage at incorrect sites (11). (*ii*) The ribozyme is able to catalyze the appropriate transesterification reaction only within a 3-base-pair window within the P1 stem. (*iii*) The precise cleavage site is selected by recognition of a wobble base pair (U·G or C·A) within this window. (*iv*) The rate of cleavage is modified by the base pair above the reactive phosphate.

The cleavage-susceptible window in P1 must be determined by factors that control the positioning of the P1 stem relative to the active site of the ribozyme. We have shown that the window is defined with reference to the bottom of the P1 stem. It is therefore likely that the bottom of the P1 stem interacts with a part of the ribozyme that is in turn loosely fixed relative to the active site. Since the identity of the bottom base pair does not seem to matter greatly, a very attractive explanation is that the P1 stem is coaxially stacked on some stem in the enzyme core. That phosphates 4, 5, and 6 in P1 can be equally susceptible to attack implies that either the ribozyme itself or the interaction of P1 with the ribozyme is quite flexible.

Within the active window on the P1 stem, catalysis occurs selectively at U·G or C·A base pairs (Fig. 4). These wobble base pairs have different substituents projecting onto both the major and minor grooves of the double helix, but a nearly identical geometry, distinct from that of the Watson–Crick base pairs or any other base–base combination. This altered base-pair geometry will necessarily lead to a locally altered sugar-phosphate backbone conformation. There are at least two ways that splice site specificity could be achieved as a result of an altered sugar-phosphate backbone conformation at the wobble position in P1. First, it is possible that the pyrimidine 3′-phosphate of a wobble base pair is positioned such that specific contacts with the intron active site are possible. A second possibility is that the wobble base pair plays a direct role in facilitating catalysis. It is likely that the reaction mechanism involves in-line attack of the guanosine 3′-hydroxyl, displacing the 5′ exon (16, 17). Inspection of A-helical RNA shows that an in-line attack would have to come from the inside of the major groove, a sterically hindered locale. A wobble base pair could result in a backbone conformation that is altered in such a way as to allow attack to come from the outside of the helix. Thus, the correct phosphate could become more accessible to attack.

Group I introns have developed an elaborate mechanism for ensuring splice-site specificity with minimal primary sequence constraint. A small site for exon–intron cleavage with most of the specificity provided by sequences within the intron could be of selective advantage to transposable self-splicing introns. These introns are almost certainly transposable elements, since they exist in many locations within individual species. The *Tetrahymena* intron has recently been shown to be capable of inserting itself into heterologous RNAs (18). If transposition involved a complementary DNA intermediate, an intron with a minimally sequence-dependent recognition site would be more likely to find insertion sites

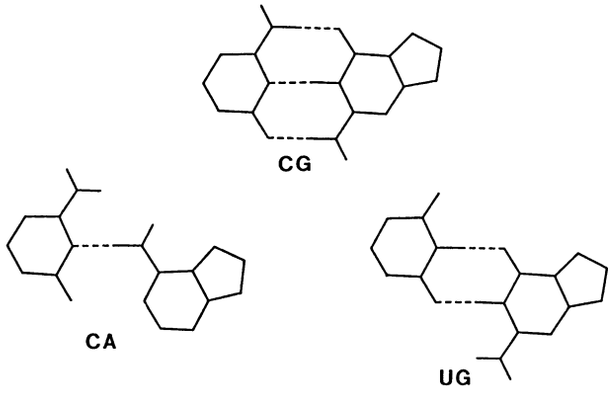

FIG. 4. Geometries of the U·G and C·A wobble base pairs, compared to the geometry of the C·G Watson–Crick base pair. Dashed lines represent hydrogen bonds.

and could therefore propagate more efficiently by transposition.

We thank A. Ellington and R. Brent for helpful discussions. This work was supported by a grant from Hoechst AG.

1. Cech, T. R., Zaug, A. J. & Grabowski, P. J. (1981) *Cell* **27,** 487–496.
2. Zaug, A. J., Grabowski, P. J. & Cech, T. R. (1983) *Nature (London)* **301,** 578–583.
3. Cech, T. R. (1987) *Science* **236,** 1532–1539.
4. Davies, R. W., Waring, R. B., Ray, J. A., Brown, T. A. & Scazzocchio, C. (1982) *Nature (London)* **300,** 719–724.
5. Michel, F. & Dujon, B. (1983) *EMBO J.* **2,** 33–38.
6. Cech, T. R., Tanner, N. K., Tinoco, I., Jr., Weir, B. R., Zuker, M. & Perlman, P. S. (1983) *Proc. Natl. Acad. Sci. USA* **80,** 3903–3907.
7. Waring, R. B., Scazzocchio, C., Brown, T. A. & Davies, R. W. (1983) *J. Mol. Biol.* **167,** 595–605.
8. Waring, R. B., Towner, P., Minter, S. J. & Davies, R. W. (1986) *Nature (London)* **321,** 133–139.
9. Been, M. D., Barfod, E. T., Burke, J. M., Price, J. V., Tanner, N. K., Zaug, A. J. & Cech, T. R. (1987) *Cold Spring Harbor Symp. Quant. Biol.* **52,** 147–157.
10. Burke, J. M., Belfort, M., Cech, T. R., Davies, R. W., Schweyen, R. J., Shub, D. A., Szostak, J. W. & Tabak, H. F. (1987) *Nucleic Acids Res.* **15,** 7217–7221.
11. Doudna, J. A., Gerber, A. S., Cherry, J. M. & Szostak, J. W. (1987) *Cold Spring Harbor Symp. Quant. Biol.* **52,** 173–180.
12. Szostak, J. W. (1986) *Nature (London)* **322,** 83–86.
13. Sanger, F., Nicklen, S. & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. USA* **74,** 5463–5467.
14. Milligan, J. F., Groebe, D. R., Witherell, G. W. & Uhlenbeck, O. C. (1987) *Nucleic Acids Res.* **15,** 8783–8798.
15. Been, M. D. & Cech, T. R. (1986) *Cell* **47,** 207–216.
16. McSwiggen, J. A. & Cech, T. R. (1989) *Science* **244,** 679–683.
17. Rajagopal, J., Doudna, J. A. & Szostak, J. W. (1989) *Science* **244,** 692–694.
18. Woodson, S. & Cech, T. R. (1989) *Cell* **57,** 335–345.