

minus the background levels observed in the HSP in the control (Sar1-GDP-containing) incubation that prevents COPII vesicle formation. In the microsome control, the level of p115-SNARE associations was less than 0.1%.

46. C. M. Carr, E. Grote, M. Munson, F. M. Hughson, P. J. Novick, *J. Cell Biol.* **146**, 333 (1999).
47. C. Ungermann, B. J. Nichols, H. R. Pelham, W. Wickner, *J. Cell Biol.* **140**, 61 (1998).
48. E. Grote and P. J. Novick, *Mol. Biol. Cell* **10**, 4149 (1999).
49. P. Uetz *et al.*, *Nature* **403**, 623 (2000).
50. GST-SNARE proteins were expressed in bacteria and purified on glutathione-Sepharose beads using standard methods. Immobilized GST-SNARE protein (0.5 μ M) was incubated with rat liver cytosol (20 mg) or purified recombinant p115 (0.5 μ M) in 1 ml of NS buffer containing 1% BSA for 2 hours at 4°C with rotation. Beads were briefly spun (3000 rpm for 10 s) and sequentially washed three times with NS buffer and three times with NS buffer supplemented with 150 mM NaCl. Bound proteins were eluted three times in 50 μ l of 50 mM tris-HCl (pH 8.5), 50 mM reduced glutathione, 150 mM NaCl, and 0.1% Triton

- X-100 for 15 min at 4°C with intermittent mixing, and elutes were pooled. Proteins were precipitated by MeOH/CH₂Cl and separated by SDS-polyacrylamide gel electrophoresis (PAGE) followed by immunoblotting using p115 mAb 13F12.
51. V. Rybin *et al.*, *Nature* **383**, 266 (1996).
52. K. G. Hardwick and H. R. Pelham, *J. Cell Biol.* **119**, 513 (1992).
53. A. P. Newman, M. E. Groesch, S. Ferro-Novick, *EMBO J.* **11**, 3609 (1992).
54. A. Spang and R. Schekman, *J. Cell Biol.* **143**, 589 (1998).
55. M. F. Rexach, M. Latterich, R. W. Schekman, *J. Cell Biol.* **126**, 1133 (1994).
56. A. Mayer and W. Wickner, *J. Cell Biol.* **136**, 307 (1997).
57. M. D. Turner, H. Plutner, W. E. Balch, *J. Biol. Chem.* **272**, 13479 (1997).
58. A. Price, D. Seals, W. Wickner, C. Ungermann, *J. Cell Biol.* **148**, 1231 (2000).
59. X. Cao and C. Barlowe, *J. Cell Biol.* **149**, 55 (2000).
60. G. G. Tall, H. Hama, D. B. DeWald, B. F. Horzodovsky, *Mol. Biol. Cell* **10**, 1873 (1999).
61. C. G. Burd, M. Peterson, C. R. Cowles, S. D. Emr, *Mol. Biol. Cell* **8**, 1089 (1997).

62. M. R. Peterson, C. G. Burd, S. D. Emr, *Curr. Biol.* **9**, 159 (1999).
63. M. G. Waters, D. O. Clary, J. E. Rothman, *J. Cell Biol.* **118**, 1015 (1992).
64. D. M. Walter, K. S. Paul, M. G. Waters, *J. Biol. Chem.* **273**, 29565 (1998).
65. N. Hui *et al.*, *Mol. Biol. Cell* **8**, 1777 (1997).
66. T. E. Kreis, *EMBO J.* **5**, 931 (1986).
67. H. Plutner, H. W. Davidson, J. Saraste, W. E. Balch, *J. Cell Biol.* **119**, 1097 (1992).
68. D. S. Nelson *et al.*, *J. Cell Biol.* **143**, 319 (1998).
69. We thank G. Waters for p115 cDNA and p115 mAbs; G. Warren for p97 and p47 antibodies; R. Scheller for rbet1, membrin, and sec22 cDNAs; H. Plutner for excellent technical assistance; and P. Tan for help during the initial phase of this work. Supported by NIH grants GM 33301 and GM42336 and National Cancer Institute grant CA58689 (W.E.B.), a NIH National Research Service Award (B.D.M.), and a Wellcome Trust International Traveling Fellowship (B.B.A.).

20 March 2000; accepted 22 May 2000

One Sequence, Two Ribozymes: Implications for the Emergence of New Ribozyme Folds

Erik A. Schultes and David P. Bartel*

We describe a single RNA sequence that can assume either of two ribozyme folds and catalyze the two respective reactions. The two ribozyme folds share no evolutionary history and are completely different, with no base pairs (and probably no hydrogen bonds) in common. Minor variants of this sequence are highly active for one or the other reaction, and can be accessed from prototype ribozymes through a series of neutral mutations. Thus, in the course of evolution, new RNA folds could arise from preexisting folds, without the need to carry inactive intermediate sequences. This raises the possibility that biological RNAs having no structural or functional similarity might share a common ancestry. Furthermore, functional and structural divergence might, in some cases, precede rather than follow gene duplication.

Related protein or RNA sequences with the same folded conformation can often perform very different biochemical functions, indicating that new biochemical functions can arise from preexisting folds. But what evolutionary mechanisms give rise to sequences with new macromolecular folds? When considering the origin of new folds, it is useful to picture, among all sequence possibilities, the distribution of sequences with a particular fold and function. This distribution can range very far in sequence space (1). For example, only seven nucleotides are strictly conserved among the group I self-splicing introns, yet secondary (and presumably tertiary) structure within the core of the ribozyme is preserved (2). Because these dis-

parate isolates have the same fold and function, it is thought that they descended from a common ancestor through a series of mutational variants that were each functional. Hence, sequence heterogeneity among divergent isolates implies the existence of paths through sequence space that have allowed neutral drift from the ancestral sequence to each isolate. The set of all possible neutral paths composes a "neutral network," connecting in sequence space those widely dispersed sequences sharing a particular fold and activity, such that any sequence on the network can potentially access very distant sequences by neutral mutations (3–5).

Theoretical analyses using algorithms for predicting RNA secondary structure have suggested that different neutral networks are interwoven and can approach each other very closely (3, 5–8). Of particular interest is whether ribozyme neutral networks approach each other so closely that they intersect. If so, a single sequence would be capable of folding into two different conformations, would

have two different catalytic activities, and could access by neutral drift every sequence on both networks. With intersecting networks, RNAs with novel structures and activities could arise from previously existing ribozymes, without the need to carry non-functional sequences as evolutionary intermediates. Here, we explore the proximity of neutral networks experimentally, at the level of RNA function. We describe a close apposition of the neutral networks for the hepatitis delta virus (HDV) self-cleaving ribozyme and the class III self-ligating ribozyme.

In choosing the two ribozymes for this investigation, an important criterion was that they share no evolutionary history that might confound the evolutionary interpretations of our results. Choosing at least one artificial ribozyme ensured independent evolutionary histories. The class III ligase is a synthetic ribozyme isolated previously from a pool of random RNA sequences (9). It joins an oligonucleotide substrate to its 5' terminus. The prototype ligase sequence (Fig. 1A) is a shortened version of the most active class III variant isolated after 10 cycles of in vitro selection and evolution. This minimal construct retains the activity of the full-length isolate (10). The HDV ribozyme carries out the site-specific self-cleavage reactions needed during the life cycle of HDV, a satellite virus of hepatitis B with a circular, single-stranded RNA genome (11). The prototype HDV construct for our study (Fig. 1B) is a shortened version of the antigenomic HDV ribozyme (12), which undergoes self-cleavage at a rate similar to that reported for other antigenomic constructs (13, 14).

The prototype class III and HDV ribozymes have no more than the 25% sequence identity expected by chance and no fortuitous structural similarities that might favor an intersection of their two neutral networks. Nevertheless, sequences can be designed that simultaneously satisfy the base-pairing requirements

Whitehead Institute for Biomedical Research and Department of Biology, Massachusetts Institute of Technology, 9 Cambridge Center, Cambridge, MA 02142, USA.

*To whom correspondence should be addressed. E-mail: dbartel@wi.mit.edu

REPORTS

of both the HDV and ligase ribozymes, while preserving most of the residues important for the activity of each ribozyme. An example of such a sequence is shown (Fig. 2A). This sequence is 42 mutational steps away from the prototype ligase (39 base substitutions, one point deletion, and two single-nucleotide insertions) and 44 mutational steps from the

prototype HDV ribozyme (40 substitutions, one deletion, and three insertions). In Fig. 2A, this sequence was color-coded in accordance with the ligase secondary structure and then threaded into the HDV secondary structure, illustrating that the two ribozyme folds have no base pairs in common. The two tertiary structures probably have no hydrogen

bonds in common, yet sequences such as this can readily be designed that might lie within, or very near to, the neutral networks for both ribozymes.

When this sequence was synthesized in the two formats of Fig. 2A, catalytic activity was detectable for both self-ligation and site-specific self-cleavage (Fig. 2B). Ligation occurred with formation of a 2',5'-phosphodiester linkage (15), the regiospecificity of the class III ligase (9), indicating that the class III ligase fold was assumed by some of the molecules. Cleavage occurred with formation of a cyclic phosphate (15), as expected for the HDV fold (11, 16). The cleavage rate was 70 times faster than the rate of uncatalyzed RNA cleavage (17, 18). The ligation rate exceeded those of nonenzymatic ligation reactions in which analogous oligoribonucleotides are aligned by simple Watson-Crick pairing. Ligation was 460 times faster than nonenzymatic formation of 2',5' linkages and seven times faster than that of 3',5' linkages (17, 19, 20). Although ligation and cleavage are slower than for the prototype ribozymes, the single sequence does catalyze both ribozyme reactions.

The effects of point mutations on this sequence provided further evidence that its position in sequence space lies at a junction of the two ribozyme folds. Substituting C13 with A simultaneously restores a residue known to be critical for ligase activity and introduces a G:A mismatch in the HDV fold (Fig. 2A, construct LIG1). The C13A point substitution dramati-

Fig. 1. Secondary structures of the prototype ribozymes. When designing ribozyme derivatives, we strived to maintain important Watson-Crick pairing (thick dashes), wobble pairing (oval), and the identity of key residues (pink letters). Secondary-structure landmarks are labeled as follows: P, paired region; J, joining segment; L, loop. (A) The class III self-ligating ribozyme. The substrate for the prototype ligase hybridizes to the ribozyme, forming the P1 helix and positioning the 2'-hydroxyl for ligation. The L3 and L5 tetraloops (UUUU) replace nonessential sequences deleted from the original class III isolate [isolate e3, (9)]. Important residues and pairing interactions were deduced by comparative analysis of active ligases that had been selected from large pools of class III variants (9). (B) The HDV antigenomic ribozyme cleaves itself at the indicated linkage (arrowhead). This sequence differs from typical antigenomic constructs in that the distal end of the P4 helix, known to be dispensable, was truncated by eight nucleotides to generate a construct 86 nucleotides long, one nucleotide shorter than the reacted ligase prototype. Important residues and pairing have been deduced from comparative analysis of HDV ribozymes and site-directed mutagenesis experiments (11). This information was rationalized and extended with the recent crystal structure of the genomic ribozyme (36).

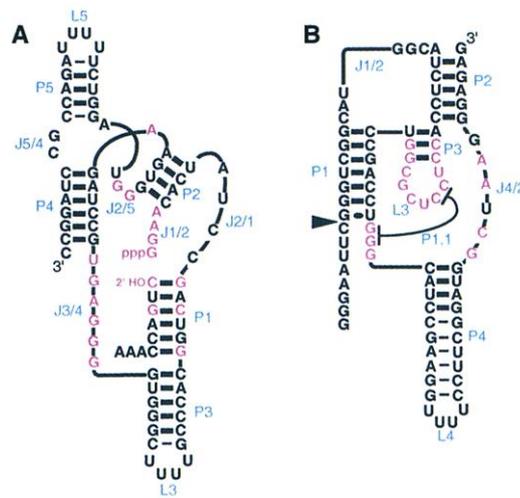
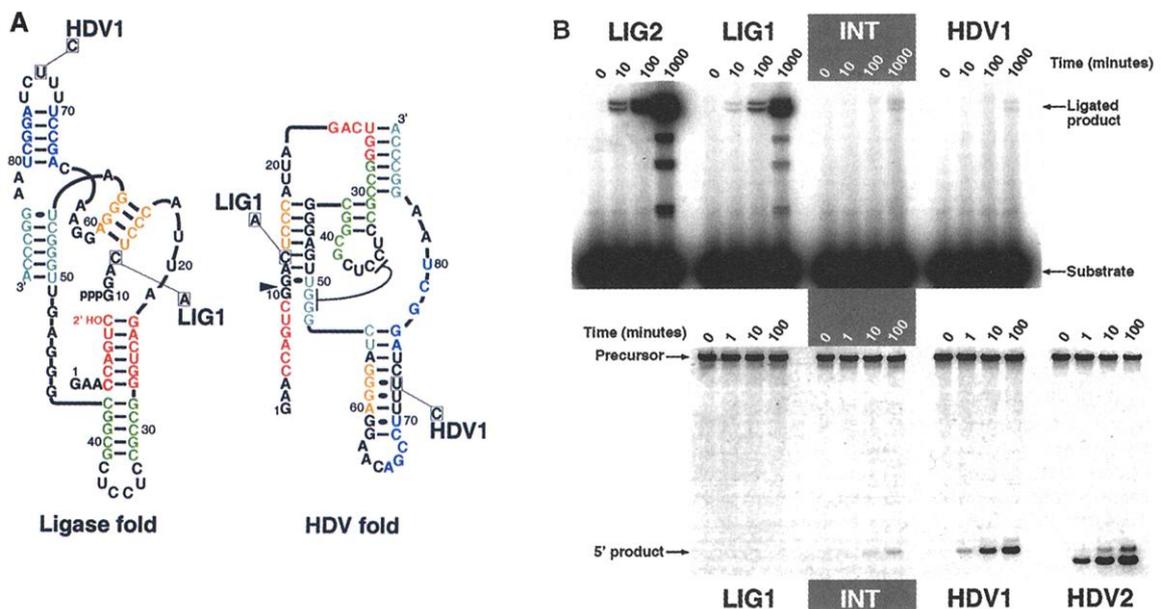


Fig. 2. One sequence, two ribozymes. (A) An intersection sequence is shown threaded through both the class III ligase and HDV ribozyme secondary structures. Colored segments are base-paired in the ligase fold. The single-nucleotide substitutions that generate constructs LIG1 and HDV1 are indicated. (B) Ligation and cleavage reactions of the intersection sequence (INT) and neighboring sequences, measured under identical conditions. LIG1 and HDV1 have single-nucleotide substitutions designed to favor ligation function and the HDV fold, respectively, as shown in (A). Substituting C38 with G in LIG1 generates LIG2, whereas deleting G10 in HDV1 generates HDV2. The top gel resolves the substrate (5'-radiolabeled) and product of ligation (70). The bottom gel resolves the precursor RNA (5'-radiolabeled) and the 5'-cleavage product (14). For LIG2, LIG1, INT, and HDV1, the ligation products are represented by two bands because of an unanticipated ligation activity of truncated ribozyme molecules (25) in addition to the activity of



full-length RNA. For these constructs, the lower product band was excluded when calculating ligation rates. For all other ligase constructs, only the full-length product band was observed. Some HDV transcripts have an extra, untemplated G at their 5' terminus, explaining the appearance of the second, slower migrating cleavage product. The deletion of G10 accounts for the faster mobility of the 5'-cleavage products of HDV2 compared with those of the intersection and HDV1 ribozymes.

ically increased the ligation rate (90 times) and lowered the cleavage rate below detection (Fig. 2B). The U73C substitution, which is expected to stabilize the HDV fold, substantially increased site-specific cleavage (120 times), while lowering the ligation rate twofold (Fig. 2, HDV1). Hence, one sequence can assume the fold of either of two ribozymes and catalyze the two respective reactions. Although the activities of this sequence are too low to conclude that the intersection of ribozyme networks can be neutral, we call it an "intersection sequence" because it demonstrates that ribozyme networks do intersect in a qualitative manner.

Ribozyme activity provides a very sensitive readout for RNA folding, but it cannot provide quantitative information on the number of sequences populating each fold at a given time. A low activity could be explained either by a small fraction of molecules populating a ri-

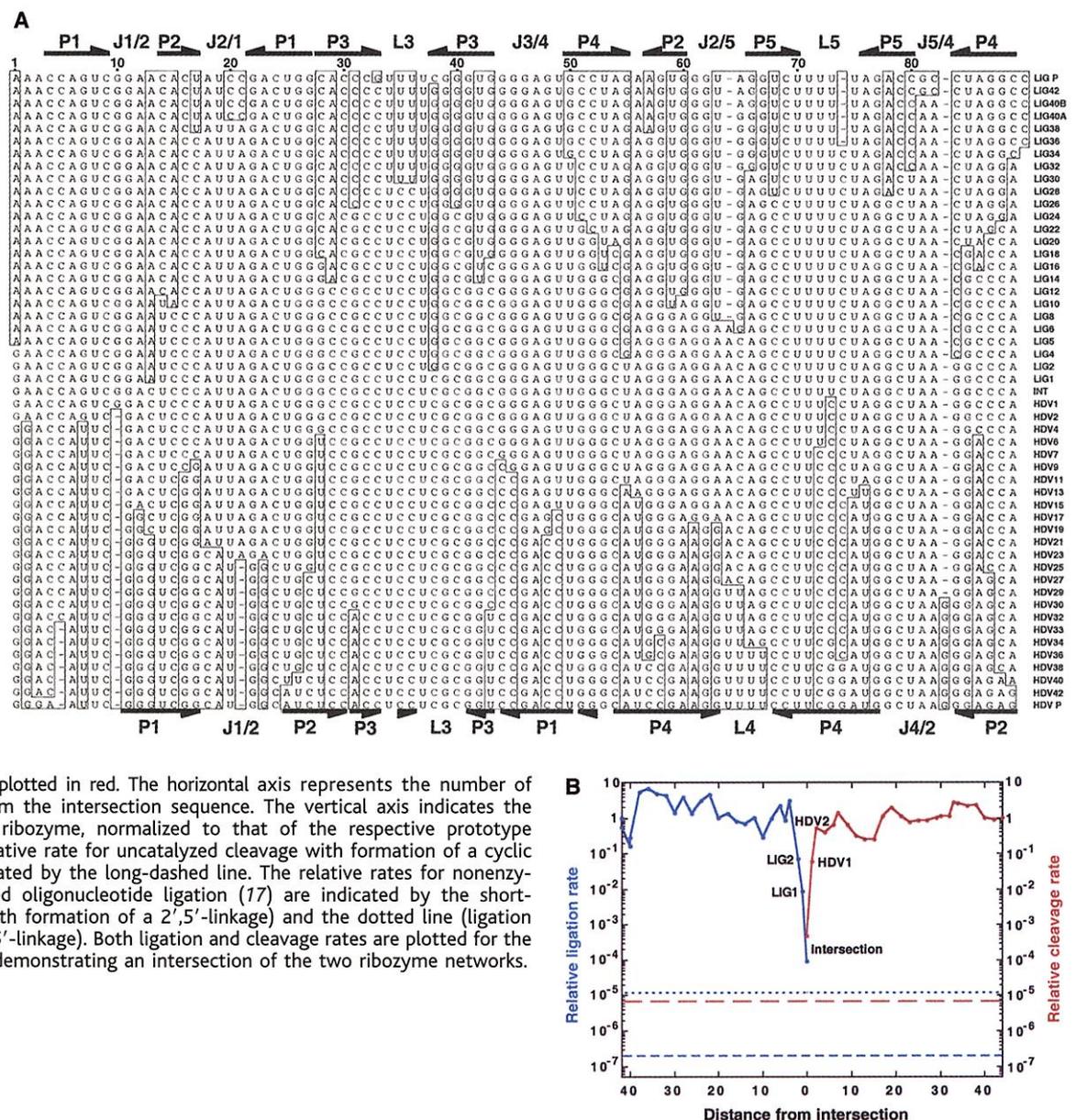
bozyme fold or by a suboptimal active site within molecules that do fold properly. Structural probing of the prototype, LIG1, intersection, and HDV1 RNAs with the Pb(II)-cleavage assay (21) indicated that both potential explanations are relevant. The dominant fold of the intersection sequence resembles that of the ligase, suggesting that suboptimal residues (e.g., C13) in the ribozyme core attenuate ligase activity. The converse is true for the HDV fold; far fewer molecules appear to assume the HDV fold, yet those that do populate the fold are highly active (22).

The substantial improvement seen with single-nucleotide substitutions suggested that the intersection sequence might be very close to both neutral networks. We searched for additional substitutions that would bring the level of activity within range of the prototype sequences (23). Substitution C38G, when combined with

the A13C substitution (construct LIG2), brought the ligation rate within 15-fold of the rate of the prototype ligase, well inside the range of variability seen for isolates of the same ribozyme from different biological species (24). The intersection sequence is also very close to the HDV neutral network; combining the G10 point deletion with the U73C substitution (construct HDV2) brought the self-cleavage rate within twofold of the prototype's rate. Thus, only four substitutions separate LIG2 and HDV2, yet the two ribozyme sequences have totally different folds and near-prototype activities.

To confirm that both LIG2 and HDV2 are on the neutral network of their respective prototypes, we designed neutral paths in sequence space that link these minor variants of the intersection sequence to their prototype sequences (Fig. 3) (25). Each step along these paths

Fig. 3. A close apposition of two ribozyme neutral networks. (A) Alignment of sequences spanning the distance between the two prototype ribozymes (25). Each sequence differs from its neighbors at no more than two residues. Each variant is named on the basis of whether it catalyzes ligation (LIG) or self-cleavage (HDV) and the number of residues that differ from the intersection sequence (boxed residues). The prototype ligase (LIG P) and HDV (HDV P) sequences are at the top and bottom of the alignment, respectively, with their secondary structures annotated as in Fig. 1. Positions are numbered with respect to the intersection sequence (INT), as in Fig. 2A. (B) Activities of the ribozyme sequences aligned in (A). Self-ligation activity is plotted in blue; self-cleavage activity is plotted in red. The horizontal axis represents the number of residues that differ from the intersection sequence. The vertical axis indicates the reaction rate of each ribozyme, normalized to that of the respective prototype ribozymes (37). The relative rate for uncatalyzed cleavage with formation of a cyclic phosphate (17) is indicated by the long-dashed line. The relative rates for nonenzymatic, template-directed oligonucleotide ligation (17) are indicated by the short-dashed line (ligation with formation of a 2',5'-linkage) and the dotted line (ligation with formation of a 3',5'-linkage). Both ligation and cleavage rates are plotted for the intersection sequence, demonstrating an intersection of the two ribozyme networks.



changed no more than two residues, often as compensatory mutations. A few of the sequences designed during this effort had less than a tenth of the prototype activity; such sequences were discarded, and alternative steps were chosen that remained on a neutral path. Very smooth paths could be designed, gradually changing nearly half the ribozyme residues yet never deviating from the prototype activities by more than sevenfold (Fig. 3B). The ease by which we were able to design these neutral sequences reinforces the idea that there are a great many neutral paths for each ribozyme fold. The ability to find neutral paths to very distant sequences, with no more than two residues changed per step, supports the notion that each fold is indeed represented by a single, contiguous neutral network rather than by multiple, isolated networks (6, 7, 26). Because these two ribozymes share no evolutionary history or structural features, it is reasonable to expect that neutral networks for other pairs of ribozymes closely approach each other. For ribozymes of similar size, it appears plausible that each ribozyme neutral network might closely approach a great many, if not all, other ribozyme networks (27).

The intersection of two networks need not be neutral to provide a viable mechanism for accessing new ribozyme folds, particularly if neutral networks very closely approach each other, as seen for the ligase and HDV ribozymes (Fig. 3). Once neutral drift generates sequences for the preexisting activity that are near to sequences with a different fold having advantageous activity, the chance mutation to an intersection sequence could impart a net selective advantage by virtue of its new biochemical function, even if the preexisting activity is greatly diminished. At this point, the gene for the intersection sequence could duplicate, and the two genes could diverge by adaptive evolution to more optimal activities, as each is relieved of the constraint imposed by providing dual activities. Thus, with intersecting networks, functional and structural divergence can precede gene duplication. This differs from the canonical view of divergence following gene duplication (28).

The likelihood that this mechanism for acquiring new folds has played an important role in the emergence of new ribozymes is difficult to ascertain. It depends on the abundance of intersection sequences that link neutral networks with biologically relevant functions and the extent to which neutral networks are explored during the course of evolution. The competing mechanism, in which new folds emerge from nonfunctional, arbitrary sequences carried as "junk" in the genome, also relies on the chance occurrence of a series of rare events. Its likelihood depends on the abundance of useful ribozymes among arbitrary sequences and the abundance of nonfunctional sequences carried in the genome. Where there is a high cost of maintaining nonfunctional sequences (for ex-

ample, in genomes of viruses, prokaryotes, and perhaps organisms of the RNA world), the intersection of ribozyme networks might provide an important mechanism for the emergence of new ribozyme folds. This could occur through a global alternative conformation, as is seen for the ligase and HDV folds, or perhaps more likely, through a series of smaller conformational switches, each involving different domains of the ancestral ribozyme. Thus, extant RNAs with no structural or functional similarity might belong to the same genealogical lineage, though no historical information has been retained at the level of sequence, structure, or function.

The finding that RNA networks intersect raises the question of whether networks for very different protein folds might also intersect (4). Some peptide segments can assume very different folds within larger protein or ribonucleoprotein contexts (29–31). However, no protein sequence is known to autonomously assume two different enzymatic folds and catalyze the two respective reactions (32). It is questionable whether such a protein sequence could be found; indeed, it has been an accomplishment to design protein sequences that have different folds yet retain 50% sequence identity (33). The chemical diversity of the 20 amino acid subunits may restrict the conformational options of protein sequences. The 20 amino acids have characteristic propensities to form α -helical or β -sheet secondary structure. They also differ markedly in water solubility, which may explain why the most dramatic protein conformational changes, such as those of prions, result in insoluble aggregates (34). In contrast, the roles of the four RNA nucleotides in forming secondary and tertiary structure are less specialized. Hence, the lack of chemical diversity among the four RNA nucleotides, often cited as a disadvantage for developing efficient catalysis, allows for comprehensive conformational flexibility, leading to the intersections of ribozyme networks and making RNA an attractive biopolymer for the birth of new functional folds in early evolution.

References and Notes

1. H. E. Aronson, W. E. Royer Jr., W. A. Hendrickson, *Protein Sci.* **3**, 1706 (1994).
2. F. Lisacek, Y. Diaz, F. Michel, *J. Mol. Biol.* **235**, 1206 (1994).
3. P. Schuster, W. Fontana, P. F. Stadler, I. L. Hofacker, *Proc. R. Soc. London B* **255**, 279 (1994).
4. A. Babajide, I. L. Hofacker, M. J. Sippl, P. F. Stadler, *Folding Des.* **2**, 261 (1997).
5. C. Reidys, P. F. Stadler, P. Schuster, *Bull. Math. Biol.* **59**, 339 (1997).
6. M. A. Huynen, P. F. Stadler, W. Fontana, *Proc. Natl. Acad. Sci. U.S.A.* **93**, 397 (1996).
7. W. Fontana and P. Schuster, *Science* **280**, 1451 (1998).
8. C. Flamm, W. Fontana, I. L. Hofacker, P. Schuster, *RNA* **6**, 325 (2000).
9. E. H. Eklund, J. W. Szostak, D. P. Bartel, *Science* **269**, 364 (1995).
10. The self-ligation rate of the prototype ligase was $1.2 \times 10^{-3} \text{ min}^{-1}$ in 10 mM Mg^{2+} at pH 7.4 and 37°C. Self-ligation of the prototype and its derivatives was

assayed as follows: 1.0 μM gel-purified ligase RNA and 0.5 μM ^{32}P -radiolabeled substrate were incubated in H_2O at 80°C for 1 min, then at 37°C for 5 min, after which the reaction was started by adding buffer and salts (30 mM tris, pH 7.4, 10 mM MgCl_2 , 200 mM KCl). Reaction aliquots were stopped by addition of EDTA, and then product and substrate were separated on 20% sequencing gels and quantitated with a Phosphorimager. The extent of reaction fit well to the single-exponential equation $f = e^{-kt}$, where f is the fraction of substrate remaining, t is time, and k is the observed rate constant for self-ligation.

11. M. D. Been and G. S. Wickham, *Eur. J. Biochem.* **247**, 741 (1997).
12. Our prototype is based on the SA1-2 antigenomic ribozyme (13). After our experiments were well underway, the antigenomic ribozyme sequence requirements were refined to include an additional 4 nucleotides of HDV sequence beyond the 3'-terminus of our prototype construct (13).
13. A. T. Perrotta and M. D. Been, *J. Mol. Biol.* **279**, 361 (1998).
14. The self-cleavage rate of our prototype HDV construct was 0.7 min^{-1} when assayed in 10 mM Mg^{2+} at pH 7.4 and 37°C. Self-cleavage reactions were carried out as described for the ligase reactions (10), except ribozyme RNA was labeled and there was no substrate RNA. Analysis of self-cleavage reactions is complicated by the nonuniform activity of molecules in a given HDV RNA preparation (11). This problem was exacerbated when a large fraction of the ribozyme reacted during the in vitro transcription reaction, leaving the uncleaved fraction enriched in molecules with slower activity, even after purification on a denaturing gel. To reduce premature cleavage in the prototype ribozyme, we shortened transcription incubations to ≤ 15 min. Including an oligodeoxynucleotide complementary to the region of the cleavage site (5'-CCGACCCGAATCCC) further reduced the premature cleavage and increased the rates observed in the self-cleavage assay, with initial rates of the HDV prototype construct increasing sevenfold (from 0.1 min^{-1} to 0.7 min^{-1}). Even with these measures, the self-cleavage reactions contained a substantial fraction of slow-reacting molecules, which presumably included misfolded molecules. Because it was important not to discount the misfolded molecules, simple initial rates are reported with no attempt to distinguish the fast-reacting fraction.
15. The 2',5' linkage was detected by exploiting the dramatically increased rate of hydrolysis of an RNA 2',5' linkage in the presence of a complementary RNA (20). The cyclic phosphate was detected by comparing the gel mobility of the 5'-cleavage product to those of known standards.
16. Ligation could not have occurred through reversal of the HDV reaction because that would have required the reactants to have a 2',3'-cyclic phosphate rather than a 5'-triphosphate at the ligation junction. Likewise, cleavage could not have occurred through reversal of the class III ligase reaction because this would have required the presence of inorganic pyrophosphate in the reaction, and it would have yielded a 3'-hydroxyl rather than cyclic phosphate at the cleavage site.
17. The uncatalyzed cleavage rate in our reaction conditions (10) is estimated to be $7.0 \times 10^{-7} \text{ min}^{-1}$ (18). The rates of template-directed, nonenzymatic oligonucleotide ligation are estimated to be $2.4 \times 10^{-10} \text{ min}^{-1}$ for 2',5'-linkage and $1.5 \times 10^{-8} \text{ min}^{-1}$ for 3',5'-linkage (19, 20).
18. Y. Li and R. R. Breaker, *J. Am. Chem. Soc.* **121**, 5364 (1999).
19. R. Rohatgi, D. P. Bartel, J. W. Szostak, *J. Am. Chem. Soc.* **118**, 3332 (1996).
20. ———, *J. Am. Chem. Soc.* **118**, 3340 (1996).
21. J. Ciesiolka, D. Michalowski, J. Wrzesinski, J. Krajewski, W. J. Krzyzosiak, *J. Mol. Biol.* **275**, 211 (1998).
22. The U73C substitution of construct HDV1, which is predicted to favor the HDV fold without changing the core of either ribozyme, does not lead to a change in the dominant fold, as determined by Pb(II) cleavage. This explains why HDV1 retains much of the ligase activity of the intersection sequence (Fig. 2B). The observation that HDV1 cleavage activity is within 20-fold of the prototype HDV ribozyme indicates

Group II Introns Designed to Insert into Therapeutically Relevant DNA Target Sites in Human Cells

Huatao Guo,¹ Michael Karberg,¹ Meredith Long,² J. P. Jones III,²
Bruce Sullenger,² Alan M. Lambowitz^{1*}

Mobile group II intron RNAs insert directly into DNA target sites and are then reverse-transcribed into genomic DNA by the associated intron-encoded protein. Target site recognition involves modifiable base-pairing interactions between the intron RNA and a >14-nucleotide region of the DNA target site, as well as fixed interactions between the protein and flanking regions. Here, we developed a highly efficient *Escherichia coli* genetic assay to determine detailed target site recognition rules for the *Lactococcus lactis* group II intron LL.LtrB and to select introns that insert into desired target sites. Using human immunodeficiency virus-type 1 (HIV-1) proviral DNA and the human *CCR5* gene as examples, we show that group II introns can be retargeted to insert efficiently into virtually any target DNA and that the retargeted introns retain activity in human cells. This work provides the practical basis for potential applications of targeted group II introns in genetic engineering, functional genomics, and gene therapy.

Group II introns are catalytic RNAs that function as mobile genetic elements by inserting directly into target sites in double-stranded DNA (1, 2). This mobility is mediated by a multifunctional intron-encoded protein (IEP) that has reverse transcriptase (RT), RNA splicing (maturase), and DNA endonuclease activities (2–5). After translation, the protein promotes RNA splicing, presumably by facilitating formation of the catalytically active intron RNA structure. It then remains associated with the excised intron to form a ribonucleoprotein (RNP) complex, which has DNA endonuclease/integrase activity. In homing, the major mobility pathway, the excised intron RNA in this complex reverse-splices into a specific target site in double-stranded DNA (6–8). The associated IEP then cleaves the opposite strand in the 3' exon of the DNA target, 9 or 10 nucleotides (nt) downstream of the intron insertion site, and uses the 3' end of the cleaved strand as a primer to reverse-transcribe the inserted intron RNA. The resulting cDNA copy of the intron is incorporated into the recipient DNA primarily by recombination mechanisms in yeast mitochondria (7) and by repair mechanisms in bacteria (8). Homing frequencies approach

100% for both fungal mitochondrial and bacterial introns (7, 8).

To initiate mobility, the intron-encoded RNP complex uses both its RNA and protein components to recognize specific sequences in its DNA target site (9–11). For the well-studied *Lactococcus lactis* LL.LtrB intron, the DNA target site extends from position –26 in the 5' exon (E1) to position +9 in the 3' exon (E2; positions numbered from the intron insertion site) (Fig. 1A) (11). A 14-nt region of the DNA target site (E1 –13 to E2 +1) is recognized primarily by base pairing with the intron RNA. This region includes short sequence elements denoted IBS2, IBS1, and δ' , which are complementary to intron sequences EBS2, EBS1, and δ (IBS and EBS refer to intron and exon binding sites, respectively) (Fig. 1, A and B). These same sequence elements are involved in base-pairing interactions required for RNA splicing (1). The regions of the DNA target site flanking the IBS and δ' sequences are recognized by the IEP. The protein first recognizes a small number of nucleotide residues in the distal 5' exon region (E1 –26 to –11) and appears to cause local DNA unwinding, enabling the intron to form base pairs with the IBS and δ' sequences for reverse splicing. Antisense-strand cleavage occurs after reverse splicing and requires additional interactions between the protein and 3' exon. The finding that at least a 14-nt region of the DNA target site is recognized by base pairing with the intron RNA raises the possibility that group II introns can be retargeted to recognize any 14-nt DNA sequence, juxtaposed to the fixed posi-

that the small fraction of HDV1 molecules that assume the HDV fold are highly active. Because U73C affects the fold and not the core of the ribozyme, the intersection molecules that assume the HDV fold must also be highly active.

23. For the ligase ribozyme, this search was informed by an experiment in which active ligase variants were selected from a large pool of minor variants of intersection sequences. A pool of more than 10^{15} variants was based on two sequences that were similar to that shown in Fig. 2A and also had detectable ligation and cleavage activity: 5'-Ggactccat tagactgggc cGCCTC-CTCG CGGCgggagt tGGCccaggg aggtaagccc ttctt-ggGcT AAgggccca and 5'-Ggactccat tagactgggc cGCCTCCTCG CGGCgggagt tGGCccaggg aggtaagccc ttctt-ggGcT AAgggccca (lowercase letters indicate positions mutagenized such that they each had a 10% probability of changing to one of the three other nucleotides; uppercase letters indicate positions not mutagenized because of their importance for HDV self-cleavage activity). Selection for ligation activity was performed as described (35). Pool activity reached that of the prototype ribozyme by round three, at which point the pool was cloned and sequenced. Each of the 20 isolates had a different nucleotide sequence, indicating that there are many highly active sequences in the vicinity of the intersection sequence. Sequences of active clones informed the design of LIG1, LIG2, and LIG4.
24. M. A. Tanner and T. R. Cech, *RNA* **2**, 74 (1996).
25. Supplemental data showing the predicted secondary structures of each construct (Fig. 3) and explaining the ligation activity of truncated ribozymes (Fig. 2B) are available at Science Online at www.sciencemag.org/feature/data/1050240.shl.
26. J. Maynard Smith, *Nature* **225**, 563 (1970).
27. The Intersection Theorem (5) states that there exists at least one intersection sequence for every pair of RNA secondary structures (usually as suboptimal conformations).
28. S. Ohno, *Evolution by Gene Duplication* (Springer-Verlag, New York, 1970).
29. D. L. Minor Jr. and P. S. Kim, *Nature* **380**, 730 (1996).
30. M. H. J. Cordes, N. P. Walsh, C. J. McKnight, R. T. Sauer, *Science* **284**, 325 (1999).
31. X. Ye et al., *Chem. Biol.* **6**, 657 (1999).
32. Although protein enzyme networks probably do not intersect, a single gene can give rise to alternative amino acid sequences having different folds through the use of overlapping reading frames. This mechanism for expression (and perhaps emergence) of multiple protein folds and functions appears prevalent in viruses, where there is a high cost to carrying non-functional sequences.
33. S. Dalal, S. Balasubramanian, L. Regan, *Folding Des.* **2**, R71 (1997).
34. F. E. Cohen, *J. Mol. Biol.* **293**, 313 (1999).
35. E. H. Eklund and D. P. Bartel, *Nucleic Acids Res.* **23**, 3231 (1995).
36. A. R. Ferré-D'Amaré, K. Zhou, J. A. Doudna, *Nature* **395**, 567 (1998).
37. Ribozyme activities were determined as described (10, 14). Because no antisense oligonucleotide was used to protect the cleavage site when transcribing the HDV variants, all cleavage rates are reported relative to the rate of the similarly prepared prototype, 0.1 min^{-1} (14). All except seven ribozymes were transcribed from synthetic oligonucleotide templates. The others (LIG P, LIG2, LIG1, INT, HDV1, HDV2, and HDV P) were transcribed from polymerase chain reaction (PCR)-amplified DNA that had been generated from plasmid DNA template and appropriate primers. This removed the small amount of sequence heterogeneity observed with long synthetic templates, eliminating concern that sequence heterogeneity might increase the apparent activity of inefficient ribozymes.
38. We thank W. Johnston for assistance with in vitro selection (23) and helpful discussions and N. Lau for assistance with Web fig. 1. We thank P. Kim as well as M. Lawrence, P. Unrau, M. Glasner, N. Bergman, and others in the lab for comments on the manuscript. Supported by the NSF/Alfred P. Sloan Foundation (E.A.S.) and the NIH (D.P.B.).

8 March 2000; accepted 18 May 2000

¹Institute for Cellular and Molecular Biology, Department of Chemistry and Biochemistry, and Section of Molecular Genetics and Microbiology, School of Biological Sciences, University of Texas, Austin, TX 78712, USA. ²Center for Genetic and Cellular Therapies, Department of Surgery, Duke University Medical Center, Box 2601, Durham, NC 27710, USA.

*To whom correspondence should be addressed. E-mail: lambowitz@mail.utexas.edu