nature structural & molecular biology

# Compact and ordered collapse of randomly generated RNA sequences

Erik A Schultes[1], Alexander Spasic[2], Udayan Mohanty[2] & David P Bartel[1]

**As the raw material for evolution, arbitrary RNA sequences represent the baseline for RNA structure formation and a standard to which evolved structures can be compared. Here, we set out to probe, using physical and chemical methods, the structural properties of RNAs having randomly generated oligonucleotide sequences that were of sufficient length and information content to encode complex, functional folds, yet were unbiased by either genealogical or functional constraints. Typically, these unevolved, nonfunctional RNAs had sequence-specific secondary structure configurations and compact magnesium-dependent conformational states comparable to those of evolved RNA isolates. But unlike evolved sequences, arbitrary sequences were prone to having multiple competing conformations. Thus, for RNAs the size of small ribozymes, natural selection seems necessary to achieve uniquely folding sequences, but not to account for the well-ordered secondary structures and overall compactness observed in nature.**



Although RNA structural biology has advanced enormously since the secondary structure of tRNA was first deduced[1], research has focused on homopolymers[2], short synthetic motifs[3] or evolved isolates that have well-defined, biologically active conformations (native states)[4]. In biological sequences, alternative conformational states induced by mutation or alteration in experimental conditions often have attenuated activity and are sometimes less compact[4–7]. It has therefore come to be assumed that compact and specific folding is the target of natural selection, and, by implication, RNAs are expected to have less organized structure in the absence of selection. Yet, because structural analyses of complex heteropolymer sequences have been limited to evolved sequences for which native states can be unambiguously defined, knowledge of RNA structure is limited by extremely biased sampling of possible sequences. Without information concerning the much larger set of unevolved sequences, this relationship between specific, compact folding and natural selection remains circumstantial.

Although early studies in RNA folding suggested extensive base pairing in randomly generated, single-stranded heteropolymers[8], it was not possible at that time to make systematic comparisons to biological sequences. Since then, folding algorithms have been developed and used to compare the computationally predicted secondary structures of evolved RNA sequences to those of their randomly permuted isomers. Random permutation erases the evolutionarily derived information that contributes to a particular fold and makes explicit the fundamental parameters that define the space of sequence possibilities from which evolution creates new folds and functions: namely, the number of nucleotide residues composing the sequence and the frequencies of the four bases. These computational studies have detected small but statistically significant differences in the predicted stability and conformational order of evolved sequences and their permutations[9–14]. Here, we implement the design of these computational studies *in vitro*, subjecting synthetic RNA constructs having either evolved or randomly generated (arbitrary) sequences to physical and chemical structural probing.

## RESULTS

### RNA constructs

First, a representative set of evolved, functional RNAs having previously characterized structures was assembled (**Table 1**). These four RNAs have no sequence similarity, range in length from 76 to 160 nucleotides and perform different biochemical and catalytic functions. The phenylalanine tRNA ($tRNA^{Phe}$) from *Saccharomyces cerevisiae*, the genomic form of the hepatitis delta virus self-cleaving ribozyme (HDV) and the P4-P6 domain of the *Tetrahymena thermophila* group I self-splicing intron have native confirmations determined by high-resolution X-ray crystalography[15–17]. The class III self-ligating ribozyme (ligase) has a secondary structure model constrained by extensive data from sequence comparisons and site-directed mutagenesis[18,19]. The ligase was isolated by *in vitro* evolution from a pool of synthetic random-sequence RNAs on the basis of its ability to catalyze the covalent linkage of a substrate oligonucleotide to its 5′ end[18].

From these four, we chose to use HDV as a model sequence to design a set of ten unique and unrelated random permutations[10]. These ten sequences were then synthesized *in vitro* and experimentally probed. Members of the permuted cohort are denoted p1 through p10. As a control for the effects of the high G+C content of these sequences, another cohort of arbitrary sequences was synthesized, again preserving sequence length, but in this case having an essentially

[1]Whitehead Institute for Biomedical Research, 9 Cambridge Center, Cambridge, Massachusetts 02142, USA. [2]Boston College Department of Chemistry, Eugene F. Merkert Chemistry Center, 2609 Beacon Street, Chestnut Hill, Massachusetts 02467, USA. Correspondence should be addressed to D.P.B. (dbartel@wi.mit.edu).

Received 12 July; accepted 21 September; published online 6 November 2005; doi:10.1038/nsmb1014

**Table 1 Size and base composition of evolved and unevolved RNA sequences**

| RNA | L[a] | A | C | G | U | G + C | M (kDa)[b] |
|---|---|---|---|---|---|---|---|
| Evolved | | | | | | | |
| $tRNA^{Phe}$ | 76 | 18 | 18 | 23 | 17 | 0.54 | 24.5 |
| HDV | 85 | 12 | 30 | 30 | 13 | 0.71 | 27.4[c] |
| Ligase | 87 | 17 | 22 | 28 | 20 | 0.57 | 28.0 |
| P4-P6 | 160 | 46 | 33 | 45 | 36 | 0.49 | 52.2 |
| Unevolved | | | | | | | |
| $Poly(U)_{85}$ | 85 | 0 | 0 | 0 | 85 | 0.00 | 26.0 |
| Permuted 1–10 | 85 | 12 | 30 | 30 | 13 | 0.71 | 27.9[c] |
| Isoheteropolymer 1–10 | 85 | 21 | 21 | 22 | 21 | 0.51 | 27.8 |

[a]L is the length of the RNA sequence in nucleotides, L = A + C + G + U. [b]M is the anhydrous molecular mass. [c]The 5′ triphosphate of the permuted sequences imparts a molecular mass slightly larger than that of the HDV, which has a 5′ hydroxyl.



uniform distribution of the four nucleotide bases. These control sequences are referred to as the isoheteropolymer cohort, denoted i1 through i10 (**Table 1**). The sequences of the evolved and arbitrary RNAs are presented in **Supplementary Figure 1** online. A polyuridine homopolymer, $poly(U)_{85}$, also having the same length as HDV, was prepared as a structural control having no sequence information, no Watson-Crick base pairing, minimal base-stacking interactions and therefore no stable uniquely folded secondary structure.

### Nondenaturing polyacrylamide gel electrophoresis

To probe the folded conformations of these RNAs, we first used PAGE, beginning with a denaturing gel and progressing to a series of nondenaturing gels. The nondenaturing gels had constant monovalent cations (30 mM $K^+$) and increasingly permissive magnesium concentrations (0, 1 and 10 mM). For evolved RNAs, delocalized $Mg^{2+}$ ions can augment the monovalent ions in inducing a universal conformational collapse by countering the electrostatic field of negatively charged backbone phosphate groups[6,7]. In some RNA sequences, $Mg^{2+}$ ions can bind site-specifically and are essential for biochemical activity[4]. We wanted to determine the response of arbitrary sequences to $Mg^{2+}$ ions.

On denaturing gels, all RNAs are expected to behave as random coils and to migrate as single, discrete bands with mobilities that are inversely proportional to the logarithm of their molecular mass. $Poly(U)_{85}$, the HDV and the permuted and isoheteropolymer RNAs had nearly identical mobilities on denaturing gels, as anticipated for RNAs of the same length (**Fig. 1a**). Under nondenaturing conditions, the distribution of RNA mobilities is influenced by the conformational collapse unique to each sequence, with compactly folding molecules tending to migrate faster in the gel. In the absence of both denaturant and $Mg^{2+}$, the HDV migrated 1.6 times faster than $poly(U)_{85}$ (**Fig. 1b**). Under the same conditions, the ligase underwent a comparable collapse to that of HDV, running slightly behind (consistent with its slightly larger mass), but in the presence of $Mg^{2+}$ the ligase overtook the HDV. As has previously been noted[20], conformational collapse was particularly striking with the P4-P6 domain. Even in the absence of $Mg^{2+}$, this 160-nt RNA comigrated with $poly(U)_{85}$, a molecule half its mass (**Fig. 1b**). In 10 mM $Mg^{2+}$, the P4-P6 domain overtook $poly(U)_{85}$ and ran midway between $poly(U)_{85}$ and the HDV.

Because the lengths and compositions of the sugar-phosphate backbones of the HDV and ligase were nearly identical to that of $poly(U)_{85}$, the additional degree of collapse in the evolved heteropolymers reflects the contribution of base-base interactions absent from the homopolymer. Under nondenaturing conditions, the

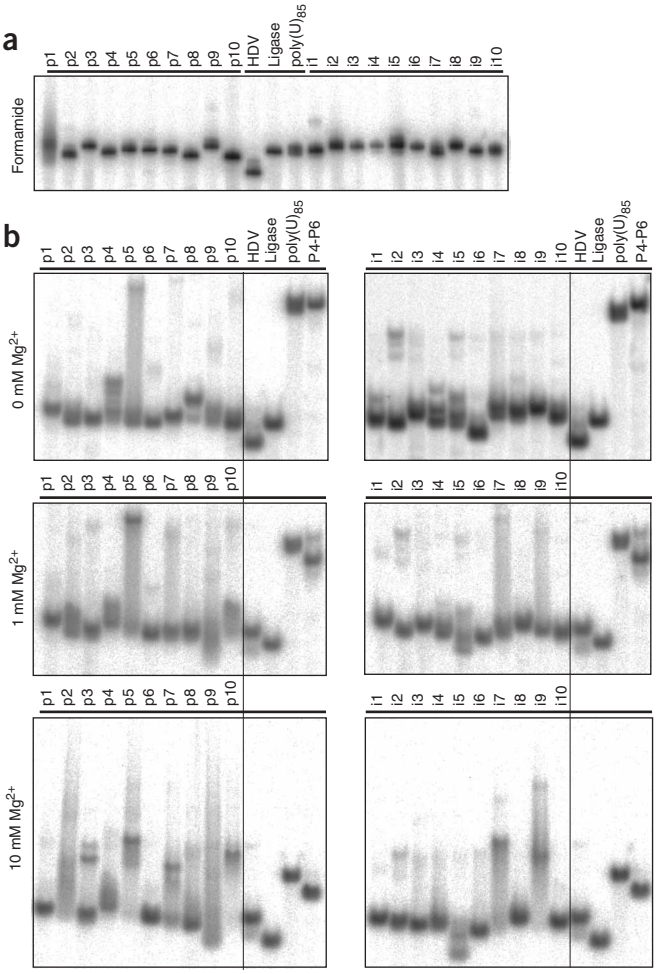


**Figure 1** PAGE of evolved and unevolved RNA sequences. (**a**) Autoradiograph of a formamide-urea denaturing gel, where mobility is a function of sequence length (molecular mass). (**b**) Autoradiographs of nondenaturing gels at three $Mg^{2+}$ concentrations, where mobility becomes a function of the shape of the folded RNA structure as well as molecular mass.

**Table 2 Minimum estimated number of alternative conformations among arbitrary sequences**

| | Number of sequences from each cohort | | | | | |
|---|---|---|---|---|---|---|
| | Permuted | | | Isoheteropolymer | | |
| $Mg^{2+}$ concentration (mM) | 0 | 1 | 10 | 0 | 1 | 10 |
| Number of major bands[a] in a lane | | | | | | |
| 1 | 8 | 5 | 4 | 7 | 8 | 8 |
| 2 | 1 | 5 | 4 | 2 | 2 | 1 |
| 3 | 0 | 0 | 2 | 1 | 0 | 0 |
| 4 | 1 | 0 | 0 | 0 | 0 | 1 |

[a]Major bands are those containing 10% or more of the total radioactivity in the lane. Each evolved sequence and poly(U)$_{85}$ had only a single major band.

arbitrary sequences of the permuted and isoheteropolymer cohorts were also highly collapsed compared to poly(U)$_{85}$. Although electrostatic collapse of arbitrary sequences is to be expected[5–7], it is notable that the degree of collapse of an arbitrary sequence was often indistinguishable from that of the HDV. Each of the 20 arbitrary sequences, in at least one of the three $Mg^{2+}$ conditions, achieved a fold that comigrated with the HDV. Subsequent chemical probing experiments (below) confirmed that this was, in most cases, the result of an ordered collapse to a specific secondary structure rather than a collapse to an amorphous, disordered state. These data suggest that compact collapsed states indistinguishable from those of evolved sequences are a common property of heteropolymeric RNA.

A single oligonucleotide sequence can often acquire multiple conformations[21]. Nondenaturing PAGE can resolve coexisting stable conformations into multiple bands in a single lane. The molar fraction acquiring each structure can be estimated by quantifying the intensity of each band (we define 'minor' bands as those having less than 10% of the total counts in the lane and 'major' bands as those having greater than 10%). Other than poly(U)$_{85}$ and the ligase, each RNA sequence underwent a unique transformation in the number of resolvable alternative folds as $Mg^{2+}$ concentration was varied. For example, the P4-P6 domain acquired a slower-moving minor band at 1 mM Mg, but not at 0 or 10 mM $Mg^{2+}$. The HDV acquired a faster-moving minor band only in the presence of $Mg^{2+}$. In contrast, one-third of the arbitrary sequences had multiple major bands (**Table 2**), and in many cases these had a smeared distribution throughout the lane, suggesting either aggregation or dynamic interconversion between two or more conformations. Subsequent sedimentation experiments (below) did not support the aggregation alternative. While the PAGE data indicated that arbitrary sequences frequently acquire compact folds, they also indicated that natural selection may be important in achieving those folds uniquely.

To compare the degree of compaction of RNA conformations from one gel to another, a model of oligonucleotide gel mobility was used to infer the effective size ($R$), in Å, of each RNA under each $Mg^{2+}$ condition (**Supplementary Methods** online). $R$ was calculated from the average mobility of the ensemble of major bands in each lane. For molecules having the same mass, smaller $R$ values corresponded to more compact folds (**Supplementary Table 1** online). Similar to poly(U)$_{85}$ and the evolved sequences, which experienced 10–17% decreases in $R$ when $Mg^{2+}$ was increased from 0 to 10 mM, 11 arbitrary sequences experienced collapses ranging from 10 to 15%. Two members of the permuted cohort (p5 and p10) and two of the isoheterpolymer cohort (i4 and i9) showed an unexpected conformational expansion with the increase in $Mg^{2+}$ concentration. Anomalous $R$ values may reflect known nonlinearities in RNA-gel interactions that can, at times, confound the simple relationship between folding, compactness and mobility. For example, single-nucleotide bulges in helical stems can create steric interactions that anomalously slow mobility[22] (artificially inflating $R$ estimates), whereas long helical stems can sometimes migrate faster than more compact, globular folds (artificially deflating $R$ estimates). Indeed, it is precisely these complexities that allow PAGE to resolve closely related conformations and provides a means to estimate the minimum number of alternative folds an individual sequence acquires *in vitro*. Another potential problem, however, in making absolute measurements of size and collapse with PAGE is that the presence of the inert gel matrix can contribute to excluded volume effects, which can promote collapse of any polymer, including RNA. Although such effects might better approximate the crowded environment of the intracellular milieu, they can complicate biophysical analyses. To control for the confounding factors of PAGE, we subjected the same set of evolved and arbitrary RNAs to a series of sedimentation experiments whereby conformational collapse can be probed uncomplicated by the vagaries of the gel phase.

## Analytical ultracentrifugation

Analytical ultracentrifugation (AUC) uses an intense centrifugal field to induce RNA sedimentation in solution, permitting an independent and complementary estimation of size[23]. Single-species sedimentation coefficients were fitted to velocity sedimentation data for each RNA at 0, 1 and 10 mM $Mg^{2+}$ (**Supplementary Table 2** online). The degree of conformational collapse for each RNA was measured by the Stokes radius ($R_s$), calculated directly from the sedimentation coefficients (Methods). $R_s$ for the evolved RNAs tended to increase with molecular mass, as expected from steric considerations alone (**Fig. 2** and **Supplementary Table 1**). For example, considering the AUC-derived

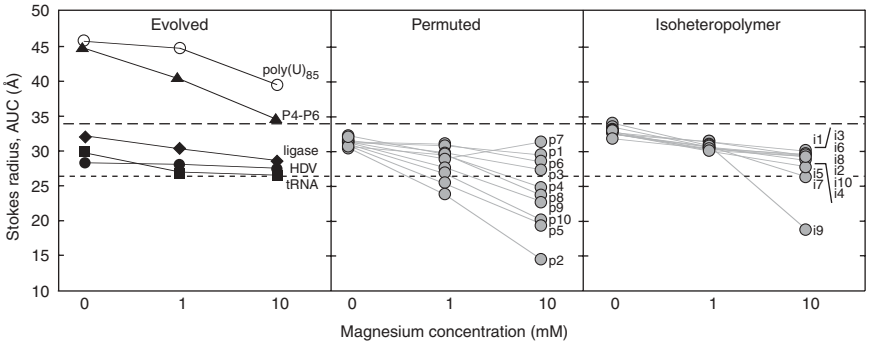


**Figure 2** Velocity sedimentation results from evolved and unevolved RNA sequences. Stokes radii of RNA conformations, as determined at three $Mg^{2+}$ concentrations, are compared. Open circle, poly(U)$_{85}$; triangle, P4-P6 domain; diamond, ligase; black filled circle, HDV; square, tRNA$^{Phe}$; gray filled circle, arbitrary sequence. Dashed lines mark the Stokes radii of HDV (short dashes) and the P4-P6 domain (long dashes) calculated from their X-ray crystal structures[40] (**Table 3**).

**Table 3 Independent estimations of Stokes radii of evolved constructs (Å)**

| RNA | $R_s$, X-ray crystal structure[a] | $R_s$, AUC experiment (10 mM $Mg^{2+}$) |
|---|---|---|
| $tRNA^{Phe}$ | 26.8 (ref. 15) | 26.2 |
| HDV | 26.1 (ref. 16) | 27.2 |
| P4-P6 | 33.5 (ref. 17) | 34.2 |

[a] $R_s$ values for the X-ray structures were calculated using HYDROPRO[40] from the atomic coordinates in their PDB files (accession codes: $tRNA^{Phe}$, 1EHZ; HDV, 1CXO; P4-P6, 1GID).



$R_s$ values at 10 mM $Mg^{2+}$, $tRNA^{Phe}$, having the least mass, also had the smallest $R_s$ (26.2 Å), followed by HDV (27.2 Å), the ligase (28.3 Å) and the P4-P6 domain (34.2 Å). These measurements were in all cases in good agreement with the corresponding X-ray crystal structures (**Table 3**). Consistent with a relatively disordered conformation, $poly(U)_{85}$ had the largest $R_s$ values (**Fig. 2**).

The AUC experiment yielded $R_s$ values that consistently decreased with increasing $Mg^{2+}$ concentration (**Fig. 2** and **Supplementary Table 2**). The only exception was p7, which had the same $R_s$ at 0 and 10 mM $Mg^{2+}$. This consistent decrease suggested that the apparent expansion of $R$ with increasing $Mg^{2+}$ observed in PAGE for four of the arbitrary sequences reflected RNA-gel interactions specific to those particular conformations and were therefore not reliable estimations of size in those cases. Similar to HDV, $tRNA^{Phe}$ and the ligase, whose respective $R_s$ values decreased by 3, 11 and 11% when $Mg^{2+}$ was increased from 0 to 10 mM, 11 arbitrary sequences experienced collapses ranging from 3 to 15% (p1, i1, i2, i3, p6, i8, i6, i10, p3, i4 and i5). The remaining 8 experienced greater $Mg^{2+}$-dependent collapse (19–53%), reminiscent of that observed for P4-P6 (**Fig. 2**).

The 11% decrease in $R_s$ observed for $tRNA^{Phe}$ upon addition of $Mg^{2+}$ (**Fig. 2**) was smaller than the 22% decrease previously observed by small-angle X-ray scattering upon transition from denaturing to native conditions[24]. In these two experiments, the sizes attained upon addition of $Mg^{2+}$ were similar to each other and matched (within 3%) the size of the native $tRNA^{Phe}$ determined by crystallography. Therefore, the two-fold difference in overall collapse is best explained by differences in the conditions before adding $Mg^{2+}$. Our conditions included 30 mM $K^+$ and lacked denaturant, thereby permitting partial folding of $tRNA^{Phe}$ in 0 mM $Mg^{2+}$. Furthermore, both the PAGE and AUC data suggest that the other evolved sequences and the arbitrary sequences also assume partially collapsed conformations before the addition of $Mg^{2+}$.

As mentioned above, in gels long helical stems can sometimes migrate faster than more compact globular folds, thereby artificially deflating $R$ estimates determined by PAGE. In AUC experiments such rod-like structures sediment more slowly than do globular folds. The observation that the arbitrary sequences generally sedimented as rapidly as did HDV indicated that the arbitrary sequences assumed configurations that were as globular as the HDV fold and that their shape did not artificially deflate $R$ estimates determined by PAGE.

Consistent with the PAGE analysis, the arbitrary sequences achieved folds substantially more compact than did $poly(U)_{85}$ (**Fig. 2**). At 0 mM $Mg^{2+}$ the $R_s$ values for the arbitrary sequences were only 2–6 Å (7–21%) larger than that determined for HDV and closely matched that of the ligase. At 1 mM $Mg^{2+}$, only 4 of the 20 arbitrary sequences had $R_s$ values that differed by more than 10% from that measured for HDV. At 1 mM $Mg^{2+}$, the extent of collapse in the permuted cohort ranged widely around that of the HDV. At 10 mM $Mg^{2+}$, the permuted cohort's collapse distribution broadened further, with p2 having an apparent collapse to only 14.1 Å, an $R_s$ approaching the theoretical limit. In contrast, the $R_s$ distribution of the isoheteropolymer cohort at 1 mM $Mg^{2+}$ remained tightly clustered, with values similar to that of the ligase and slightly above that of HDV. With the exception of i9 (18.5 Å), $R_s$ values of the isoheteropolymer cohort remained relatively constrained around those of HDV and ligase at 10 mM $Mg^{2+}$. Because the fitted masses from the sedimentation experiments were typically close approximations of the theoretical masses (within 5%; **Supplementary Table 2**), implying no detectable deviation from ideality, aggregation can be ruled out as a possible explanation accounting for most of the observed collapsed states.

At 10 mM $Mg^{2+}$, 8 of the 20 arbitrary sequences had achieved folds more compact than HDV (**Fig. 2**), with the arbitrary G+C-rich sequences (permuted cohort) tending to access more compact folds and a broader range of collapse than compositionally neutral sequences (isoheteropolymer cohort). Of these eight, six (p8, p9, p10, p5, p2 and i9) had $R_s$ values more than 2.7 Å (10%) smaller than the HDV value. Is it possible that so many arbitrary sequences could be folding substantially more compactly than HDV? Several of these sequences (p8, p10, p5 and p2) were among the minority of sequences having poor fits to the theoretical mass (>10% deviation; **Supplementary Table 2**), raising concern about the accuracy of their $R_s$ values. However, the other two (p9 and i9) had excellent agreement between their fitted and theoretical masses (differing by 1% and 2% respectively), suggesting that RNAs with folds more compact than that of HDV were more common than might have been expected. Together, the results from PAGE and AUC suggested that the compact states observed among biological RNAs reflect a generic capacity for compact folding irrespective of natural selection and, therefore, independent of the adaptive constraints that are idiosyncratic to particular biochemical functions.

## Chemical probing

PAGE and AUC analyses reveal the gross anatomical features of molecular conformations. To gain insight into the details of the collapsed states and secondary structures acquired by arbitrary sequences, we employed a chemical probing technique using divalent lead cations. Coordinated Pb(II)-bound hydroxyl groups readily remove the proton from the 2′-OH of the ribose ring and promote a nucleophilic attack on the adjacent phosphate, which cleaves the RNA backbone into fragments having 5′-OH and cyclic 2′,3′-phosphates[25]. Lead-induced cleavage is elevated at residues where the backbone is free to assume conformations permitting in-line attack. Linkages within a Watson-Crick helix are among those that are inflexible and protected from lead cleavage. Incubating structured RNA with lead acetate, therefore, results in a distribution of cleavage products that, when separated on a sequencing gel, correspond roughly to the secondary structure[26].

Chemical probing experiments at 10 mM $Mg^{2+}$ were performed on three of the evolved RNAs ($tRNA^{Phe}$, HDV and ligase), $poly(U)_{85}$ and 19 of the arbitrary sequences (p1 did not migrate as a discrete band on denaturing gels (**Fig. 1a**), making its lead-probing results uninterpretable). The degree of cleavage at each site was determined from autoradiographs by calculating the difference in intensities between the no-lead control and the probed RNA at the 90-s time point (shown for HDV in **Fig. 3a**). These differences were then assigned a cleavage coefficient ranging from 0 at protected sites (no cleavage detected above the no-lead background rate) to 3, the value that dominated the uniform cleavage profile of the random-coil $poly(U)_{85}$ (**Fig. 3a** and **Supplementary Fig. 2** online). Individual linkages that

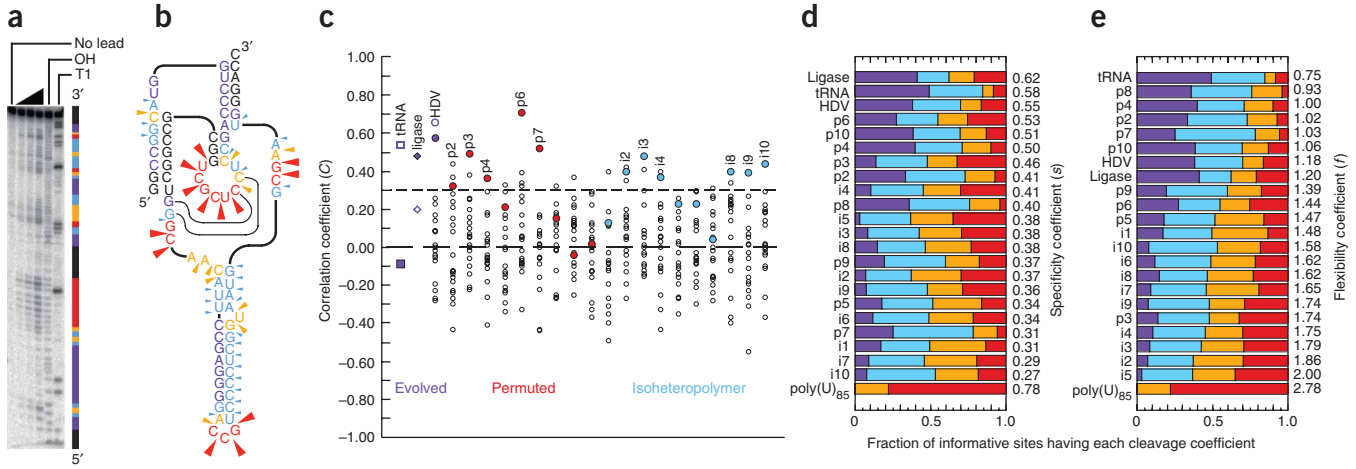


**Figure 3** Results of chemical probing of evolved and unevolved RNA sequences. (**a**) Autoradiograph resolving the lead-induced cleavage products of the HDV ribozyme. Wedge indicates increasing length of lead incubation (30, 90 and 600 s). Partial base hydrolysis (OH) and partial digestion with T1 ribonuclease (T1), which cuts after G residues, provided size markers. Colored column represents cleavage coefficients assigned to each position (purple, 0; blue, 1; orange, 2; red, 3). Reliable data could not be obtained from the extreme top and bottom of the gels or from positions near residues 27–37, where buffers interfered with RNA gel mobility (black). (**b**) Cleavage coefficients mapped onto the experimentally determined secondary structure of the HDV ribozyme. Arrows indicate by color (as in **a**) and size (larger arrows depict higher cleavage rates) the susceptibility of each link to cleavage by lead(II) ions. Each nucleotide position is also color coded to indicate cleavage of its 3′ link, a convention followed for all other sequences in **Supplementary Figure 2**. (**c**) The correlations between the observed lead-induced cleavage and secondary structures of the 22 heteropolymer RNAs. The correlations between cleavage and cognate computationally predicted secondary structure are plotted with solid symbols (purple, evolved RNAs; red, permuted cohort; blue, isoheteropolymer cohort); those between cleavage and the empirically derived cognate secondary structures of evolved RNAs are plotted with corresponding purple open symbols; and those between cleavage and the noncognate computationally predicted secondary structures derived from the 19 other sequences are plotted with small black open circles. Dashed lines mark no correlation ($C = 0.00$, long dashes) and the 5% significance threshold ($C = 0.30$, short dashes). (**d**) The fraction of informative sites having each cleavage coefficient, sorted by the specificity coefficient ($s$). Poly(U)$_{85}$ is a special case where this approximation of conformational specificity does not hold and is positioned at the bottom of the chart to maintain logical consistency. (**e**) The fraction of informative sites having each cleavage coefficient, sorted by the flexibility coefficient ($f$).

showed hyperelevated cleavage (two times above the median value of poly(U)$_{85}$; **Supplementary Fig. 2**), which are suggestive of fortuitous lead-specific binding sites[25], were not included in the analysis.

Consistent with previous lead-probing experiments[27], the lead cleavage pattern for HDV corresponded well to its known secondary structure (**Fig. 3b**). To quantitate the correspondence between the lead cleavage pattern and the secondary structures of tRNA$^{Phe}$, ligase and HDV, we first assigned to each nucleotide a linkage value corresponding to the known secondary structures of these RNAs (either 0, 3 or 1.5, depending on whether the link joined two paired residues, two unpaired residues or a paired and unpaired residue, respectively). We then calculated the correlation between these values and the lead-cleavage coefficients (**Fig. 3c**, open symbols for tRNA$^{Phe}$, ligase and HDV). Similarly, the correlations were calculated using the structures computationally predicted by MFOLD[28] (**Fig. 3c**, solid symbols for tRNA$^{Phe}$, ligase and HDV; secondary structures depicted in **Supplementary Fig. 2**). The known cloverleaf motif of tRNA$^{Phe}$ correlated well with its observed distribution of cleavage sites ($C = 0.54$), but its computationally predicted structure did not (–0.09). Conversely, in the case of the ligase, it was the computationally predicted secondary structure that was more highly correlated (0.48) than the experimentally determined pseudoknot structure (0.20), an indication that the active conformation of this ribozyme is not the most populous. HDV's cleavage profile had a high correlation with its computationally predicted structure (0.58) and a somewhat higher correlation (0.65) with its double-pseudoknotted X-ray crystal structure[16].

MFOLD[28] was then used to calculate predicted structures for the arbitrary sequences (**Supplementary Fig. 2**). Of the 19 arbitrary sequences with lead-cleavage data, 18 had positive correlation coefficients between their cleavage profiles and their predicted structures (**Fig. 3c**, filled red and blue circles). To evaluate the statistical significance of these correlations, we calculated the correlation coefficients using the predicted secondary structures of the noncognate RNAs (**Fig. 3c**, small open circles; **Supplementary Methods**). The correlation coefficients with the noncognate structures should approximate those that would be expected by chance if there were no relationship between the lead-cleavage pattern and the predicted secondary structure. For HDV and 11 of the 19 arbitrary sequences, the correlation with the cognate predicted structure was statistically significant ($P < 0.05$), and for 8 arbitrary sequences the cognate predicted structure correlated better than any of the noncognate structures (**Fig. 3c**). The significance of the correlations demonstrated that structural collapse among these arbitrary sequences was relatively ordered and that these RNAs acquired a unique sequence-dependent secondary structure or set of closely related secondary structures not unlike those of the evolved RNAs. Furthermore, these correlations were conservative estimates of ordered collapse among arbitrary sequences, because, as the example of tRNA$^{Phe}$ shows, low correlations can be obtained not from lack of ordered structure but from inaccurate structural predictions.

Having provided evidence that arbitrary sequences undergo ordered collapse, we set out to quantitatively estimate the degree to which arbitrary sequences fold into specific secondary structures. Multiple or interconverting conformations are expected to result in a preponderance of ambiguous cleavage coefficients with intermediate values (1 and 2), whereas sequences that generate unique and specific folds would have fewer intermediate cleavage coefficients. For folded sequences, the fraction of cleavage coefficients with extreme values

(0 and 3), *s*, is an approximation of the specificity of folding. The evolved sequences acquired their conformations with highest specificity (**Fig. 3d**), indicating that extrinsic constraints (imposed by natural selection) are essential for achieving folds uniquely. This was consistent with the PAGE results, where the evolved sequences showed only a single major band with few additional minor bands, whereas the arbitrary sequences frequently showed multiple major bands (**Table 2**). Nonetheless, there was no qualitative discontinuity between the specificity coefficients of the evolved and arbitrary sequences, indicating that unevolved, heteropolymeric RNAs frequently assume a secondary structure (or small number of related structures) not qualitatively different than those of biological RNAs.

The average cleavage coefficient of the informative sites in a sequence, *f*, is an approximate measure of the overall conformational flexibility of the phosphodiester backbone of folded RNA. When the 23 RNAs were rank-ordered by *f* (**Fig. 3e**), $tRNA^{Phe}$ had the least flexibility in its structure ($f = 0.75$), whereas $poly(U)_{85}$ had the most (2.78). Rank-ordering the RNAs with respect to *f* effectively sorted the two cohorts (with the exception of p3). The HDV and ligase RNAs had similar flexibility coefficients (1.18 and 1.20, respectively) that were, in turn, similar to the average flexibility of the permuted cohort (1.23) but substantially lower than the average flexibility of the isoheteropolymer cohort (1.71). Although it might be expected that G+C-rich sequences would tend to have more rigidly constrained backbones than compositionally neutral sequences, the precision with which the flexibility coefficient seemed to distinguish between the two cohorts was notable. Along with the systematic differences in conformational collapse between the two arbitrary sequence cohorts observed using AUC (**Fig. 2**), these data suggest that base composition can, for certain biophysical properties, have more impact on RNA structure than do the details of the nucleotide sequence.

Because of the effects of base composition on the compactness, specificity and flexibility of RNA structure, the relative contribution of intrinsic and extrinsic constraints shaping evolved conformations must be a function of the base composition of the evolving nucleotide sequence, as has been proposed from theoretical studies[29]. For example, the lead-cleavage results indicate that the low flexibility of the HDV is accounted for exclusively by the physiochemical constraints intrinsic to its G+C-rich sequence. Yet, the $tRNA^{Phe}$ and ligase sequences, having nearly uniform base compositions, had lower flexibilities than those of the isoheteropolymer cohort (**Fig. 3e**). The conformational properties of the $tRNA^{Phe}$ and ligase RNAs, therefore, imply the existence of extrinsic constraints imposed by natural selection via the nonrandomness of their nucleotide sequences. For those RNA heteropolymers that have base compositions yielding more poorly ordered conformations, natural selection has to 'search harder' to find those relatively rare sequences having sufficiently compact, specific and rigid folds.

## DISCUSSION

Our results indicate that arbitrary sequences frequently assume compact folds and specific secondary structures, implying that natural selection need not be invoked to explain the occurrence of these properties in biological RNAs. However, we cannot speak to the degree to which arbitrary sequences assume well-defined tertiary structure because the analytical methods used here cannot be used to interpret the nature of tertiary-level intramolecular interactions. The well-defined secondary structure elements within the arbitrary sequences could acquire one or a few stable tertiary folds or, if linked by flexible hinge-like regions, they could remain in a highly compact but dynamic state of interconverting tertiary conformations. It is reasonable to expect the latter—that stabilization of specific tertiary structure is crafted by the extrinsic forces of natural selection. By analogy to protein folding, the self-consistency principle described in ref. 30 dictates that short-range interactions alone can only constrain folding to a distribution of local conformations that is typically broader than that observed in native states. It is the long-range interactions (that is, contacts between nucleotide residues distantly separated in the sequence) that are necessary for the formation of specific tertiary structure. An important example in RNA is the ubiquitous A-minor motif, in which an unpaired adenosine residue can dock into the minor groove of helical elements, forming long-range interactions shown to be necessary for defined tertiary folding and function[31,32]. However, because the long-range contacts needed to stabilize specific tertiary folds are presumably more demanding on the details of the sequence than are the Watson-Crick paring combinations forming secondary structure, specific tertiary conformations are probably relatively rare absent natural selection.

Arbitrary protein sequences rarely fold into well-ordered structures[33]. As a further complication, only about 20% of protein sequences generated randomly with respect to the 20 amino acids are soluble[34]. When compared to protein, RNA seems to be intrinsically more soluble and also more ordered—at least to the extent that the secondary structure detected in the arbitrary RNAs indicates ordered folding (**Fig. 3c**). To help compensate for the intrinsic difficulty in obtaining stable, soluble protein folds, biologically informed patterns in amino acid composition have been used to focus the search for protein structure or function to provinces of sequence space where those properties are more common[35–37]. Understanding the effects of G+C content on the specificity and flexibility of RNA structure might inform the design of compositionally focused combinatorial libraries for selection of functional RNAs, although it seems that RNA has much less to gain from such strategies than does protein.

The physical and chemical probing data described here define more precisely the relationship between biochemical function, natural selection and folding. For RNAs the size of small ribozymes, physiochemical properties intrinsic to the four nucleotide bases, the ribose ring and the phosphodiester backbone ensure that defined and compact secondary structures are ubiquitous in sequence space. Random mutational processes will, therefore, readily access RNAs with sequence-specific secondary structures regardless of selection pressures. At the same time, these physiochemical properties also permit individual sequences to acquire multiple secondary structure conformations, with perhaps many more tertiary conformations. An evolutionary consequence of these intrinsic properties is the great phenotypic variation that would be accessible to sequences under limited mutational searches. If any of the alternative conformations of a sequence should have fortuitous functionality, differential survival and reproduction could then operate to select those sequence variants that prefer the conformation having the new function. The spontaneous generation among arbitrary sequences of a limited set of multiple folds that are both compact and soluble would thus have made RNA the ideal substrate for the emergence of structurally dependent biochemical function early in the history of life.

## METHODS

**RNA preparation, nondenaturing gel electrophoresis, analytical ultracentrifugation, chemical probing and statistical analysis of chemical probing data.** See **Supplementary Methods.**

**RNA renaturation.** All RNAs were renatured by heating in water to 80 °C for 2 min, equilibrating to 22 °C for 5 min and then adding buffer and salts to final

concentrations of 66 mM Tris-HEPES (pH 7.5), 0.5 mM EDTA, 30 mM KCl and either 0, 1 or 10 mM $MgCl_2$. RNAs were equilibrated in buffer and salts for at least 20 min before being subjected to structural analysis.

**Calculating compactness of folded RNA.** Quantitative estimates of the degree of compactness of RNA conformations were obtained independently from PAGE mobility and AUC sedimentation data. For PAGE results, a suitable generalization of the screened hydrodynamic formalism[38] was adopted that describes the absolute electrophoretic mobility of an ionic oligomer in polyacrylamide gel and that models the macromolecule as having rod-like elements that assume an overall spherical conformation with an effective radius $R$ (**Supplementary Methods**). In calculating the compactness of the folded RNA from AUC, the Stokes radius, $R_s$, was computed directly from the definition of the sedimentation coefficient, $s$:

$$s = m(1-\bar{\nu}\rho)/(6\pi\eta R_s)$$

where $s$ is derived from the experiment; $m = M/N$, where $M$ is the molecular mass of the RNA and $N$ is Avogadro's number; $\bar{\nu}$ is the specific volume of RNA (conventionally accepted as 0.53 $cm^3\ g^{-1}$); $\rho$ is the density of buffer (1.01 g $cm^{-3}$); $\eta$ is the viscosity coefficient ($9.548 \times 10^{-3}$) (ref. 39).

*Note: Supplementary information is available on the Nature Structural & Molecular Biology website.*

ACKNOWLEDGMENTS

We thank U. Müller, J.G. Ruby, M.S. Lawrence and J.S. Philo for comments on the manuscript; P. Hraber and M. Deras for computational assistance; V. Carey for statistical consultation; and R. Burton, G. Hersch and R. Sauer for use of the XL-A ultracentrifuge. This work was supported by grants from the US National Institutes of Health to D.P.B, from the Medical Foundation to E.A.S and from the US National Science Foundation to A.S. and U.M.

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

Published online at http://www.nature.com/nsmb/
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/



1. Holley, R.W. *et al.* Structure of a ribonucleic acid. *Science* **147**, 1462–1465 (1965).
2. Fresco, J.R. From DNA melting profiles to tRNA crystals and RNA chaperones. in *RNA Structure and Function* (ed. Grunberg-Manago, M.) 1–35 (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, USA, 1998).
3. Xia, T. *et al.* Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. *Biochemistry* **37**, 14719–14735 (1998).
4. Batey, R.T., Rambo, R.P. & Doudna, J.A. Tertiary motifs in RNA structure and folding. *Angew. Chem. Int. Edn. Engl.* **38**, 2326–2343 (1999).
5. Buchmueller, K.L. & Weeks, K.M. Near native structure in an RNA collapsed state. *Biochemistry* **42**, 13869–13878 (2003).
6. Heilman-Miller, S.L., Thirumalai, D. & Woodson, S.A. Role of counterion condensation in folding of the *Tetrahymena* ribozyme I. Equilibrium stabilization by cations. *J. Mol. Biol.* **306**, 1157–1166 (2001).
7. Heilman-Miller, S.L., Pan, J., Thirumalai, D. & Woodson, S.A. Role of counterion condensation in folding of the *Tetrahymena* ribozyme II. Counterion-dependence of folding kinetics. *J. Mol. Biol.* **309**, 57–68 (2001).
8. Doty, P., Boedtker, J.R., Fresco, J.R. & Haselkorn, M.L. Secondary structure in ribonucleic acids. *Proc. Natl. Acad. Sci. USA* **45**, 482–499 (1959).
9. Fresco, J.R., Alberts, B.M. & Doty, P. Some molecular details of the secondary structure of ribonucleic acid. *Nature* **188**, 98–101 (1960).
10. Schultes, E.A., Hraber, P.T. & LaBean, T.H. Estimating the contributions of selection and self-organization in RNA secondary structure. *J. Mol. Evol.* **49**, 76–83 (1999).
11. Seffens, W. & Digby, D. mRNAs have greater negative folding free energies than shuffled or codon choice randomized sequences. *Nucleic Acids Res.* **27**, 1578–1584 (1999).
12. Higgs, P.G. RNA secondary structure: Physical and computational aspects. *Q. Rev. Biophys.* **33**, 199–253 (2000).
13. Le, S.Y., Zhang, K.Z. & Maizel, J.V. RNA molecules with structure dependent functions are uniquely folded. *Nucleic Acids Res.* **30**, 3574–3582 (2002).
14. Clote, P., Ferre, F., Kranakis, E. & Krizanc, D. Structural RNA has lower folding energy than random RNA of the same dinucleotide frequency. *RNA* **11**, 578–591 (2005).
15. Shi, H.J. & Moore, P.B. The crystal structure of yeast phenylalanine tRNA at 1.93 angstrom resolution: A classic structure revisited. *RNA* **6**, 1091–1105 (2000).
16. Ferre-D'Amare, A.R., Zhou, K. & Doudna, J.A. Crystal structure of a hepatitis delta virus ribozyme. *Nature* **395**, 567–574 (1998).
17. Cate, J.H. *et al.* Crystal structure of a group I ribozyme domain: Principles of RNA packing. *Science* **273**, 1678–1685 (1996).
18. Ekland, E.H., Szostak, J.W. & Bartel, D.P. Structurally complex and highly-active RNA ligases derived from random RNA sequences. *Science* **269**, 364–370 (1995).
19. Schultes, E.A. & Bartel, D.P. One sequence, two ribozymes: Implications for the emergence of new ribozyme folds. *Science* **289**, 448–452 (2000).
20. Juneau, K. & Cech, T.R. *In vitro* selection of RNAs with increased tertiary structure stability. *RNA* **5**, 1119–1129 (1999).
21. Uhlenbeck, O.C. Keeping RNA happy. *RNA* **1**, 4–6 (1995).
22. Lilley, D.M. Analysis of global conformation of branched RNA species using electrophoresis and fluorescence. *Methods Enzymol.* **317**, 368–393 (2000).
23. Philo, J.S. An improved function for fitting sedimentation velocity data for low-molecular-weight solutes. *Biophys. J.* **72**, 435–444 (1997).
24. Fang, X. *et al.* $Mg^{2+}$-dependent compaction and folding of yeast $tRNA^{Phe}$ and the catalytic domain of the *B. subtilis* RNase P RNA determined by small-angle X-ray scattering. *Biochemistry* **39**, 11107–11113 (2000).
25. Brown, R.S., Dewen, J.C. & Klug, A. Crystallographic and biochemical investigation of the lead(II)-catalyzed hydrolysis of the yeast phenylalanine tRNA. *Biochemistry* **24**, 4785–4801 (1985).
26. Krzyzosiak, W.J. *et al.* Characterization of the lead(II)-induced cleavages in tRNAs in solution and effect of the Y-base removal in yeast $tRNA^{Phe}$. *Biochemistry* **27**, 5771–5777 (1988).
27. Matysiak, M., Wrzesinski, J. & Ciesiolka, J. Sequential folding of the genomic ribozyme of the hepatitis delta virus: Structural analysis of RNA transcription intermediates. *J. Mol. Biol.* **291**, 283–294 (1999).
28. Mathews, D.H., Sabina, J., Zucker, M. & Turner, H. Expanded sequence dependence of thermodynamic parameters provides robust prediction of RNA secondary structure. *J. Mol. Biol.* **288**, 911–940 (1999).
29. Schultes, E.A., LaBean, T.H. & Hraber, P.T. A parameterization of RNA sequence space. *Complexity* **4**, 61–71 (1999).
30. Go, N. Theoretical studies of protein folding. *Annu. Rev. Biophys. Bioeng.* **12**, 183–210 (1983).
31. Nissen, P., Ippolito, J.A., Ban, N., Moore, P.B. & Steitz, T.A. RNA tertiary interactions in the large ribosomal subunit: The A-minor motif. *Proc. Natl. Acad. Sci. USA* **98**, 4899–4903 (2001).
32. Battle, D.J. & Doudna, J.A. Specificity of RNA-RNA helix recognition. *Proc. Natl. Acad. Sci. USA* **99**, 11676–11681 (2002).
33. Hecht, M.H., Das, A., Go, A., Bradley, L.H. & Wei, Y.N. *De novo* proteins from designed combinatorial libraries. *Protein Sci.* **13**, 1711–1723 (2004).
34. Prijambada, I.D. *et al.* Solubility of artificial proteins with random sequences. *FEBS Lett.* **382**, 21–25 (1996).
35. Davidson, A.R., Liumb, K.J. & Sauer, R.T. Cooperatively folded proteins in random sequence libraries. *Nat. Struct. Biol.* **2**, 856–864 (1995).
36. Davidson, A.R. & Sauer, R.T. Folded proteins occur frequently in libraries of random amino acid sequences. *Proc. Natl. Acad. Sci. USA* **91**, 2146–2150 (1994).
37. Wei, Y.N. & Hecht, M.H. Enzyme-like proteins from an unselected library of designed amino acid sequences. *Protein Eng. Des. Sel.* **17**, 67–75 (2004).
38. Mohanty, U. & McLaughlin, L.W. On the characteristics of migration of oligomeric DNA in polyacrylamide gels and in free solution. *Annu. Rev. Phys. Chem.* **52**, 93–106 (2001).
39. Bloomfield, V., Crothers, D.A. & Tinoco, I. *Physical Chemistry of Nucleic Acids* (Harper and Row, New York, USA, 1974).
40. Fernandes, M.X., Ortega, A., Lopez Martinez, M.C. & Garcia de la Torre, J. Calculation of hydrodynamic properties of small nucleic acids from their atomic structure. *Nucleic Acids Res.* **30**, 1782–1788 (2002).