Global similarities in nucleotide base composition among disparate functional classes of single-stranded RNA imply adaptive evolutionary convergence

ERIK SCHULTES,1,2 PETER T. HRABER,3 and THOMAS H. LABEAN2,4
1Department of Earth and Space Sciences, University of California at Los Angeles, Los Angeles, California 90024, USA
2Combinatorial Sciences Center, Duke University Medical Center, Durham, North Carolina 27710, USA
3Department of Biology, University of New Mexico, Albuquerque, New Mexico 87131, USA
4Department of Biochemistry, Duke University Medical Center, Durham, North Carolina 27710, USA

ABSTRACT
The number of distinct functional classes of single-stranded RNAs (ssRNAs) and the number of sequences representing them are substantial and continue to increase. Organizing this data in an evolutionary context is essential, yet traditional comparative sequence analyses require that homologous sites can be identified. This prevents comparative analysis between sequences of different functional classes that share no site-to-site sequence similarity. Analysis within a single evolutionary lineage also limits evolutionary inference because shared ancestry confounds properties of molecular structure and function that are historically contingent with those that are imposed for biophysical reasons. Here, we apply a method of comparative analysis to ssRNAs that is not restricted to homologous sequences, and therefore enables comparison between distantly related or unrelated sequences, minimizing the effects of shared ancestry. This method is based on statistical similarities in nucleotide base composition among different functional classes of ssRNAs. In order to denote base composition unambiguously, we have calculated the fraction G+A and G+U content, in addition to the more commonly used fraction G+C content. These three parameters define RNA composition space, which we have visualized using interactive graphics software. We have examined the distribution of nucleotide composition from 15 distinct functional classes of ssRNAs from organisms spanning the universal phylogenetic tree and artificial ribozymes evolved in vitro. Surprisingly, these distributions are biased consistently in G+A and G+U content, both within and between functional classes, regardless of the more variable G+C content. Additionally, an analysis of the base composition of secondary structural elements indicates that paired and unpaired nucleotides, known to have different evolutionary rates, also have significantly different compositional biases. These universal compositional biases observed among ssRNAs sharing little or no sequence similarity suggest, contrary to current understanding, that base composition biases constitute a convergent adaptation among a wide variety of molecular functions.

Keywords: composition space; RNA evolution; RNA simplex

INTRODUCTION
The remarkably successful comparative sequence analyses of ribosomal RNA have had as their main goals the elucidation of rRNA structure and the construction of phylogenetic relationships between divergent sequences (Woese et al., 1990a; Woese & Pace, 1993). However, this requires the existence of well conserved, homologous sites amiable to comparison (Pace et al., 1989) and therefore imposes two important limitations on evolutionary inference. First, it is impossible to define the relationships between sequences having no sequence similarity (e.g., comparing 23S rRNA with P RNA) (Gould, 1991). Second, restricting sequence analyses to single evolutionary lineages (i.e., homologous sequences), makes impossible the characterization of molecular properties that are universal to functional ssRNAs. This is because shared genealogies confound similarities due to biophysical constraints with similarities that are the result of historically contingent factors (Harvey & Pagel, 1991; Kauffman, 1993). Although comparative sequence analyses can resolve secondary structure effectively within a functional class, they severely limit generalized inferences about structural

Reprint requests to: Erik Schultes, Center for the Study of the Evolution and Origin of Life, Geology Building, University of California, Los Angeles, Los Angeles, California 90024, USA; e-mail: schultes@abacus.mc.duke.edu.
properties and biophysical mechanisms (e.g., the RNA folding problem) that may be common to RNA polymers in general. Hence, research in molecular evolution has been directed largely toward historical reconstructions of the branching order of sequence diversification and the analysis of structural adaptations specific to individual lineages. Studies addressing the generic properties of evolved RNA polymers and the origin of new function are relatively rare (Bloch et al., 1983, 1985; Fontana et al., 1993; Tomizawa, 1993; Huynen & Hogeweg, 1994). In an attempt to avoid these problems, we have developed a method of comparative analysis of ssRNA utilizing statistical distributions of attributes that are common to any RNA sequence, even those sharing no evolutionary history (i.e., having no common ancestor). The main goal of this statistical approach is not the construction of phylogenetic relationships, but establishing and explaining global similarities between sequences that are known to be distantly related or completely unrelated.

In this paper, we compare the statistical distributions of nucleotide base composition of different functional classes of ssRNAs. Analyses of nucleic acid base composition have typically focused on G+C content data derived from the thermal denaturation/hybridization or density gradient sedimentation experiments using large genomic DNA samples (Chargaff et al., 1949; Chargaff, 1951; Sabeur et al., 1993). This study differs in two respects. First, we examine the full range of nucleotide composition calculated as the fraction of A, C, G, and U residues in a sequence of N residues (A/N, C/N, G/N, U/N). This quantity is referred to as a composition vector, and can be calculated for any RNA sequence. Although commonly used by itself, the percent G+C content is a one-dimensional projection of (4-1)-dimensional data [i.e., knowing A/N, C/N, and G/N implies the value of U/N because (A/N + C/N + G/N + U/N) = 1.0]. The compression of complex composition information onto the single dimension of G+C content results in the loss of information. Because these four fractions must sum to one, all possible composition vectors can be visualized as points within the volume of a tetrahedron. This geometric representation is called a unit simplex and constitutes the entire composition space of all possible RNA sequences. We have visualized this space (Fig. 1) using interactive graphics software developed by Richardson and Richardson (1992). The composition vectors describing the four homopolymers (poly-A, poly-C, etc.) are represented by the vertices, whereas the point located at the center, equidistant from the four homopolymers, represents all those sequences having a uniform distribution of nucleotides, i.e., (0.25, 0.25, 0.25, 0.25). We refer to this special class of heteropolymers as the isoheteropolymers. In general, an arbitrary composition class "contains" a large number of possible sequences, many sharing no sequence similarity. Although composition vectors specify base composition exactly, they can be cumbersome to work with analytically. Taking the fraction G+C content as a convention, we invoke a more convenient notation using two additional measures of base composition; the fraction G+A and fraction G+U content. Together, G+C, G+A, and G+U measures act as a coordinate system, uniquely locating position in RNA composition space.

This study also differs from previous base composition analyses in that we narrow our focus from bulk genomic samples (containing a wide variety of regulatory, coding, and noncoding DNA sequences) to distinct ssRNA sequences that are known to play specific metabolic roles in the cell. Similar to proteins, ssRNA molecules fold in a sequence-specific manner into complex conformations that determine their metabolic function (Draper, 1992, 1996; Price & Nagai, 1996; Zarrinkar & Williamson, 1996). By focusing our analysis on disparate classes of ssRNA sequences, it may be possible to correlate properties in base composition to biochemical properties that are characteristic of functional ssRNAs, such as intramolecular folding and structure. Also, the lack of sequence similarity between functional classes implies either independent origins (Ekland et al., 1995) or radical sequence divergence early in evolutionary history, before the latest common an-
The ancestor of contemporary life (Lewin, 1985; Woese & Pace, 1993). Comparing disparate sequences minimizes the confounding influences of genealogy on the interpretation of molecular attributes as either historical accident or as generic properties of RNA polymers. We have compiled and plotted within the RNA simplex, composition vectors of 2,800 complete sequences from 14 distinct and naturally occurring functional classes of ssRNA molecules and three classes of artificial ribozymes. Within each functional class, except tRNA, each genus is represented by a single randomly chosen species, making it representative of the universal phylogenetic tree. This database represents a diverse cross section of RNA composition (G+C content ranging from 0.088, mitochondrial GTC-tRNA Drosophila yakuba, to 0.748, 5S rRNA, Thermomicrobium roseum); structure (sequence length varies from 13 nt, a modified hepatitis delta virus ribozyme a, to 5,182 nt, 23S rRNA Homo sapiens); function (including various ribozymes); organisms and ecological settings (including extremeophiles such as Halobacterium, Thermus, and Pyrodicium); and includes artificial, self-ligating ribozymes evolved in vitro (Bartel & Szostak, 1993) (Table 1). The structural requirements of full-length messenger RNAs are less well understood and, as such, mRNAs lie beyond the scope of this analysis.

Relative to the well-known variability in G+C content, we have discovered among the empirical distributions an unexpected, universal localization of base composition in G+A and G+U content. We present four general observations: (1) the empirical distributions form a set of parallel axes in composition space that are themselves parallel to the gradient in G+C content; (2) these axes are displaced in a magnitude and direction indicating persistent and relatively constrained G+A and G+U biases; (3) the variance of the G+A and G+U biases are dependent on sequence length; and (4) secondary structural elements of unrelated sequences nonetheless have characteristically similar base composition biases; stems are G+C- and G+U-rich, whereas loops are G+C-poor and G+A-rich. The universal nature of these base composition biases between disparate sequences unrelated by structure or function is evidence that base composition in itself has a universal adaptive value independent of nucleotide sequence and invites a reexamination of the role base composition plays in the folding dynamics, function, and evolution of RNA polymers.

RESULTS

Comprehensive overview

Taken as a whole, the mean of the G+C, G+A, and G+U content of the 2,800 natural ssRNA sequences investigated here are each slightly greater than 0.5 (Table 1). The mean composition vector of these data is (0.242, 0.235, 0.274, 0.249), guanosine residues predominating, G+A is the least variable of the composition measures, followed by G+U. The variability of the base composition is greatest in G+C, being 6.5 times more variable than G+A and 4.3 times more variable than G+U. Below we partition these data by functional class and phylogenetic domain.

Empirical distributions parallel Chargaff's Axis

Chargaff established for genomic DNA that the mole-
fraction of A is equal to T and C is equal to G (Chargaff et al., 1949; Chargaff, 1951). In the RNA simplex, those composition vectors fulfilling the equivalent of Chargaff's Rule (A = U and simultaneously C = G) form a line joining the midpoints of the AU and CG edges and contain the isoheteropolymer composition vector. We refer to this locus of composition vectors as Chargaff's Axis (Fig. 1). In almost every case studied here (exceptions being four classes of small nuclear RNA), the distributions of ssRNA composition vectors are noticeably protruded into axes that lie parallel to Chargaff's Axis (Fig. 2, panels labeled i). These empirical axes, extending from G+C-rich to G+C-poor compositions, are another way of visualizing the well-known variability in G+C content between species. For the majority of the distributions, the mean G+C content is greater than 0.5, and is exceptionally high among the Archaea. Particularly striking is the distinct G+C biases between mitochondrial and chloroplast tRNA sequences (Fig. 2). Chloroplast sequences are slightly G+C-rich, whereas mitochondrial sequences show remarkably low G+C content with 91% of the 742 mitochondrial tRNAs having G+C values less than 0.5.

A standard model for the diversification of RNA sequences is the compensatory change among paired nucleotides. This is the basis of structural inference from sequence comparison studies. Thus, AU pairs can often replace CG pairs and vice versa with only neutral effects (Pace et al., 1989). Compensatory mutations allow ssRNAs to drift neutrally and therefore relatively rapidly within functional extremes of low to high G+C content (Wada, 1992), without affecting overall G+A and G+U content (Fig. 4). Hence, evolutionary diversification driven by compensatory mutations will result in axis-like distributions of neutral mutations (or "neutral ridges," Schuster et al., 1994) lying parallel to but displaced from Chargaff's Axis, as seen in panels labeled ii of Figure 2. G+C biases can be driven by mechanisms that are external to the organism [e.g., extreme environmental parameters such as temperature or salinity tending to increase G+C content (Brown et al., 1993)] or internal to the organism [e.g., cytosine deamination mutation pressures or polymerase error biases tending to decrease G+C content (Pearl & Savva, 1996)]. We examine these issues in the Discussion.
# TABLE 1. Phylogenetically and functionally representative compilation of ssRNA nucleotide composition.

<table>
<thead>
<tr>
<th>RNA</th>
<th>Taxon</th>
<th>n*</th>
<th>(N)*</th>
<th>(G+C)*</th>
<th>(G+U)*</th>
<th>(G+A)*</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Comprehensive</td>
<td>2,800</td>
<td>287.0 ± 662.0</td>
<td>0.509 ± 0.200</td>
<td>0.516 ± 0.0310</td>
<td>0.523 ± 0.0462</td>
</tr>
<tr>
<td></td>
<td>23S rRNA</td>
<td>15</td>
<td>2,968.9 ± 65.8</td>
<td>0.588 ± 0.0525</td>
<td>0.567 ± 0.00597</td>
<td>0.509 ± 0.0212</td>
</tr>
<tr>
<td></td>
<td>Bacteria</td>
<td>39</td>
<td>2,915.6 ± 86.3</td>
<td>0.526 ± 0.0383</td>
<td>0.570 ± 0.00720</td>
<td>0.517 ± 0.0102</td>
</tr>
<tr>
<td></td>
<td>Eucarya</td>
<td>33</td>
<td>3,615.0 ± 470.3</td>
<td>0.530 ± 0.0832</td>
<td>0.540 ± 0.0139</td>
<td>0.520 ± 0.0143</td>
</tr>
<tr>
<td></td>
<td>Metazoa</td>
<td>20</td>
<td>1,821.0 ± 43.2</td>
<td>0.494 ± 0.0297</td>
<td>0.517 ± 0.0123</td>
<td>0.535 ± 0.0113</td>
</tr>
<tr>
<td></td>
<td>16S rRNA</td>
<td>15</td>
<td>1,530.7 ± 184.9</td>
<td>0.611 ± 0.0427</td>
<td>0.562 ± 0.00500</td>
<td>0.507 ± 0.00681</td>
</tr>
<tr>
<td></td>
<td>Bacteria</td>
<td>85</td>
<td>1,511.8 ± 50.9</td>
<td>0.550 ± 0.0387</td>
<td>0.568 ± 0.00770</td>
<td>0.520 ± 0.0118</td>
</tr>
<tr>
<td></td>
<td>Eucarya</td>
<td>47</td>
<td>1,823.6 ± 57.8</td>
<td>0.486 ± 0.0255</td>
<td>0.524 ± 0.00671</td>
<td>0.527 ± 0.00955</td>
</tr>
<tr>
<td></td>
<td>Aracha</td>
<td>26</td>
<td>124.5 ± 3.9</td>
<td>0.598 ± 0.0726</td>
<td>0.508 ± 0.0272</td>
<td>0.498 ± 0.0409</td>
</tr>
<tr>
<td></td>
<td>Bacteria</td>
<td>123</td>
<td>117.6 ± 4.9</td>
<td>0.575 ± 0.0574</td>
<td>0.520 ± 0.0308</td>
<td>0.495 ± 0.0344</td>
</tr>
<tr>
<td></td>
<td>Eucarya</td>
<td>234</td>
<td>119.3 ± 1.4</td>
<td>0.557 ± 0.0369</td>
<td>0.517 ± 0.0119</td>
<td>0.505 ± 0.0176</td>
</tr>
<tr>
<td></td>
<td>Aracha</td>
<td>7</td>
<td>400.9 ± 62.0</td>
<td>0.644 ± 0.0966</td>
<td>0.564 ± 0.0140</td>
<td>0.469 ± 0.0296</td>
</tr>
<tr>
<td></td>
<td>Bacteria</td>
<td>37</td>
<td>389.0 ± 39.7</td>
<td>0.569 ± 0.114</td>
<td>0.570 ± 0.0168</td>
<td>0.469 ± 0.0147</td>
</tr>
<tr>
<td></td>
<td>Aracha</td>
<td>13</td>
<td>757.15 ± 517.5</td>
<td>0.434 ± 0.105</td>
<td>0.561 ± 0.0302</td>
<td>0.492 ± 0.0268</td>
</tr>
<tr>
<td></td>
<td>Bacteria</td>
<td>6</td>
<td>734.8 ± 853.8</td>
<td>0.355 ± 0.0933</td>
<td>0.566 ± 0.0259</td>
<td>0.495 ± 0.0276</td>
</tr>
<tr>
<td></td>
<td>Aracha</td>
<td>2</td>
<td>43.5 ± 43.1</td>
<td>0.619 ± 0.0212</td>
<td>0.512 ± 0.0368</td>
<td>0.519 ± 0.0269</td>
</tr>
<tr>
<td></td>
<td>Bacteria</td>
<td>24</td>
<td>162.4 ± 4.2</td>
<td>0.557 ± 0.0228</td>
<td>0.487 ± 0.00957</td>
<td>0.548 ± 0.0152</td>
</tr>
<tr>
<td></td>
<td>Aracha</td>
<td>16</td>
<td>187.8 ± 11.6</td>
<td>0.455 ± 0.0357</td>
<td>0.456 ± 0.0310</td>
<td>0.543 ± 0.0257</td>
</tr>
<tr>
<td></td>
<td>Bacteria</td>
<td>8</td>
<td>220.6 ± 14.6</td>
<td>0.468 ± 0.0684</td>
<td>0.477 ± 0.0417</td>
<td>0.561 ± 0.0207</td>
</tr>
<tr>
<td></td>
<td>Eucarya</td>
<td>11</td>
<td>143.6 ± 12.3</td>
<td>0.484 ± 0.0384</td>
<td>0.493 ± 0.0192</td>
<td>0.536 ± 0.0303</td>
</tr>
<tr>
<td></td>
<td>Aracha</td>
<td>11</td>
<td>125.6 ± 29.7</td>
<td>0.411 ± 0.0345</td>
<td>0.448 ± 0.0279</td>
<td>0.531 ± 0.0220</td>
</tr>
<tr>
<td></td>
<td>Bacteria</td>
<td>14</td>
<td>101.5 ± 7.7</td>
<td>0.451 ± 0.0233</td>
<td>0.544 ± 0.0304</td>
<td>0.469 ± 0.0352</td>
</tr>
<tr>
<td></td>
<td>Aracha</td>
<td>2011</td>
<td>74.3 ± 6.3</td>
<td>0.496 ± 0.120</td>
<td>0.510 ± 0.0280</td>
<td>0.528 ± 0.0509</td>
</tr>
<tr>
<td></td>
<td>Bacteria</td>
<td>121</td>
<td>77.0 ± 4.2</td>
<td>0.633 ± 0.0457</td>
<td>0.503 ± 0.0215</td>
<td>0.516 ± 0.0351</td>
</tr>
<tr>
<td></td>
<td>Eucarya</td>
<td>371</td>
<td>78.3 ± 5.2</td>
<td>0.580 ± 0.0522</td>
<td>0.506 ± 0.0222</td>
<td>0.522 ± 0.0317</td>
</tr>
<tr>
<td></td>
<td>Aracha</td>
<td>436</td>
<td>75.8 ± 4.2</td>
<td>0.572 ± 0.0438</td>
<td>0.510 ± 0.0282</td>
<td>0.544 ± 0.0365</td>
</tr>
<tr>
<td></td>
<td>Bacteria</td>
<td>291</td>
<td>75.6 ± 5.0</td>
<td>0.526 ± 0.0531</td>
<td>0.510 ± 0.0212</td>
<td>0.540 ± 0.0341</td>
</tr>
<tr>
<td></td>
<td>Mitochondria</td>
<td>742</td>
<td>70.3 ± 6.4</td>
<td>0.372 ± 0.0939</td>
<td>0.514 ± 0.0328</td>
<td>0.519 ± 0.0685</td>
</tr>
<tr>
<td></td>
<td>Aracha</td>
<td>1</td>
<td>274</td>
<td>0.471</td>
<td>0.511</td>
<td>0.493</td>
</tr>
<tr>
<td></td>
<td>Bacteria</td>
<td>1</td>
<td>119</td>
<td>0.513</td>
<td>0.546</td>
<td>0.445</td>
</tr>
<tr>
<td></td>
<td>Eucarya</td>
<td>1</td>
<td>271</td>
<td>0.568</td>
<td>0.535</td>
<td>0.546</td>
</tr>
<tr>
<td></td>
<td>Mitochondria</td>
<td>1</td>
<td>273</td>
<td>0.487</td>
<td>0.502</td>
<td>0.524</td>
</tr>
<tr>
<td></td>
<td>Aracha</td>
<td>1</td>
<td>273</td>
<td>0.546</td>
<td>0.524</td>
<td>0.502</td>
</tr>
<tr>
<td></td>
<td>Bacteria</td>
<td>1</td>
<td>274</td>
<td>0.544</td>
<td>0.522</td>
<td>0.518</td>
</tr>
</tbody>
</table>

n* is the number of individual sequences in the data set.
N is the length of the sequences in nucleotide residues. ( ) notation indicates mean value.
G+C, G+U, and G+U were calculated as the fraction of these nucleotides in individual sequences. ( ) notation indicates mean value.
Variability is specified as ± one SD.

The random-sequence domains of these molecules are 220 nt, however, the full-length sequences (including the constant primer binding regions) were used in calculating the base composition. This is justified because the constant regions are present during selection and could conceivably participate in the self-ligation reaction. Constant region compositions are: G+C (5' primer 0.591; 3' primer 0.350), G+A (5' primer 0.500; 3' primer 0.400), and G+U (5' primer 0.364; 3' primer 0.500). The G+A bias among the random-sequence domains (averaging 0.340 ± 0.014) accounts for the overall bias observed among the full-length artificial sequence data.
Universal G+A and G+U biases

If one views the RNA simplex from the C = G = 0.5 endpoint of Chargaff's Axis, one sees Chargaff's Axis and the empirical axes mentioned above, end-on (Fig. 2, panels labeled ii). In this perspective, it is clear that the variability of distributions in directions perpendicular to the G+C gradient are relatively constrained and, with the exception of five classes of snRNA, consistently G+A-rich. Unlike G+C content, which is variable, the G+A and G+U content is unexpectedly constant within and between functional classes, particularly for the longer sequences. The location of these axes with respect to Chargaff's Axis can be quantified by calculating the means and standard deviations in G+A and G+U (Table 1; Fig. 3).

For functional classes containing longer sequences (23S, 18S, 16S, P RNA) common to each phylogenetic domain (Archaea, Bacteria, and Eucarya), we observe that composition vectors from each domain cluster into individual axis, each being distinct from Chargaff's Axis. Intriguingly, the phylogenetic arrangement of the Archaea, Bacteria, and Eucarya distributions is remarkably similar between functional classes, having nearly identical mean G+A and G+U values [compare, for example, mean G+A values for Archaea 23S rRNA (0.567), 16S rRNA (0.562), and P RNA (0.564) in Table 1]. These domain-specific G+A and G+U biases between different functional classes of ssRNAs suggest that characteristic constraints act on the evolution of base composition within domains. The G+A and G+U content of shorter sequences, such as 5S rRNA and tRNA, are less systematically biased than longer sequences and are more variable. We describe this length dependence in detail in the next section.

Dependence of G+A and G+U variability on sequence length

The variance of G+A and G+U biases in the empirical distributions are dependent on the length of the RNA sequences (Fig. 5). As populations of RNA sequences diversify through mutation and selection, they map a system of branching trajectories through sequence space. Changes in sequence composition along these trajectories can be observed in the RNA simplex. The simplest trajectories are those of totally neutral, random walks, where sequences accept single point mutations at some constant rate, uniformly over the four nucleotides. This "ergodic" process tends to drive trajectories toward the isoheteropolymer composition. However, for sequences of biologically relevant length, random trajectories can be quite variable. This is because single base substitutions in short sequences have a larger relative effect on nucleotide frequency than in longer sequences. Hence, the distributions of shorter RNA sequences (tRNA, snRNA, 5S rRNA) are "buffered" more easily by point mutations than longer sequences (e.g., 23S rRNA, P RNA). This length effect can be seen in distributions of randomly generated sequences that form spherical clouds centered on the isoheteropolymers (G+C = (G+A) = (G+U) = 0.5 where "< >" notation refers to mean values) (Fig. 2). As sequence length increases, the variance about the mean values of G+A, G+C, and G+U content decrease isotropically. The variability in the composition of these random-sequence distributions accounts well for the observed changes in variability in G+A and G+U content of naturally evolved ssRNAs as a function of mean sequence length. Sequences having fewer than roughly 200 nt are an exception, having smaller variability than expected. The decrease in the range and variability of the distributions with increasing sequence length tends to sharpen inherent phylogenetic distinctions between the Archaea, Bacteria, and Eucarya axes. However, unlike the random distributions, the mean G+A and G+U biases do not converge to 0.5 with increasing sequence length (Fig. 6). This gives the impression that longer sequences better reflect intrinsic G+A and G+U bias than do shorter sequences. Because of the effects of compensatory mutation, G+C content re-
FIGURE 3. Distributions of mean G+A and G+U contents of six functional classes of ssRNA sequences. The location of the long axis of the data distributions with respect to Chargaff's Axis are described using the mean G+A and G+U content of each domain and functional class. Bars represent one SD. This plot is simply an analytical projection of the simplex depicted in Figure 2, panels labeled ii, where the red lines are represented here by the vertical (G+U gradient) and horizontal (G+A gradient). Panel F depicts seven self-ligating ribozymes isolated from in vitro evolution of synthetic, random-sequence libraries. These ribozymes enhance the rate of the self-ligation reaction by more than $10^6$ times the rate of the uncatalyzed reaction. Class I, Class II, and Class III ribozymes are evolved independently (sharing no common ancestor), yet they share similar base compositions among themselves and with natural sequences.
Global similarities among disparate ssRNA

FIGURE 4. Evolutionary diversification of ssRNA via compensatory mutations results in the observed axes lying parallel to but displaced from Chargaff’s Axis. **A:** Rate of nucleotide substitution in stems and loops of *E. coli* 16S rRNA. From phylogenetic comparisons, Neefs et al. (1993) calculated the rate of base substitution at each site in the 16S rRNA sequence and grouped them into six categories from low to high variability. Using the inferred secondary structure for this molecule, we calculated the number of sites in both paired and unpaired regions for each category of substitution rate. Loop regions are dominated by sites having low to moderate substitution rates, whereas the stem regions are dominated by sites having high rates of substitution. These data indicate that stems have nearly a four times higher rate of evolution than loops. This high rate of mutation in stem regions is the result of compensatory changes in Watson-Crick partners, and is thought to have relatively small effects on structure and function, and are therefore relatively neutral. **B:** Hypothetical “neutral ridges” in RNA composition space. The ridge (blue) is derived by calculating the base composition of *E. coli* RNase P RNA as if the stems had sustained compensatory mutations of increasing or decreasing G+C contents. The wild-type P RNA is depicted in yellow. The ridge is an axis, parallel to Chargaff’s Axis, and displaced by a magnitude and direction stipulated primarily by the composition of the loop regions.

mains extremely variable and independent of sequence length.

**Base composition of secondary structural elements**

RNA secondary structure can be decomposed into base paired stem regions and unpaired loop regions. Assuming Watson–Crick base pairing only, the base composition of the stem regions must be such that A = U and C = G, and therefore will be limited to Chargaff’s Axis [i.e., (G+A) = (G+U) = 0.5, regardless of G+C]. If this assumption holds, any compositional bias taking the observed composition vector off Chargaff’s Axis would therefore be due to biases in the loop regions. Using sequence data with inferred secondary structures, we have calculated the base composition of both stems and loops (Table 2; Fig. 7). Similar to results from other studies (Woese et al., 1990b; Vawter & Brown, 1993; Gutell et al., 1994; Cate et al., 1996), we document (1) a slight deviation from Chargaff compositions in the stem regions and (2) the compositional biases between paired and unpaired bases are significantly different from one another.

Stems are slightly G+A- and G+U-rich due to non-canonical G-A and G-U interactions (approximately 2% and 7% of the total base pair interactions, respectively). Stems are also heavily biased in G+C content, having an average 17% higher G+C content than the average of the full-length sequence (mean stem G+C content ± one SD is 0.667 ± 0.0555). In contrast, the loops are purine-rich (mean loop G+A = 0.605 ± 0.0693) and remarkably G+C-poor (mean loop G+C = 0.378 ± 0.0571). It is primarily the purine bias of the loop regions that yield the observed, overall purine bias in these sequences. The nonuniform base composition between different secondary structural elements suggests the existence of differential selection pressures biasing nucleotide substitution rates with respect to base composition.

**DISCUSSION**

**Factors affecting the evolution of base composition**

The observed localization of disparate ssRNAs in G+A and G+U content suggests (1) G+A and G+U content is subject to severe evolutionary constraints and (2) these constraints are functionally or phylogenetically universal. We refer to this confined volume of composition space occupied by disparate ssRNA as an evolutionary attractor in the sense that evolutionary trajectories of diverse and disparate sequences have remained invariant in base composition despite drastic evolutionary divergence at the sequence level. The documentation of an evolutionary attractor implies a radically different view of molecular evolution.
tionally, differences among related sequences are used to reconstruct putative phylogenetic branching histories as an individual lineage undergoes diversification. In contrast, we have discovered similarities among unrelated sequences that need not necessarily be explained by phylogenetic constraints. The evolutionary mechanisms accounting for an attractor in composition space can be classified into three broad categories: mutation biases, historical constraints imposed by ancestral sequences (Sueoka, 1992), and selection due to general physiochemical constraints on RNA structure. These factors are outlined below.

Mutation

The fundamental driving force behind evolutionary change is the creation of new variation via mutation. Mutational mechanisms in DNA replication and repair that are not compositionally uniform might be expected to impose biases on the coding regions of ssRNA (Pearl & Savva, 1996). For example, spontaneous cytosine deamination to uracil in genomic DNA creates an ever persistent mutation "pressure" toward uracil residues, which, if not recognized and repaired by appropriate pathways, become fixed as thymine residues. Consequently, G·C pairs mutate to A·T pairs, decreasing the G+C content of genomic DNA and presumably ssRNA coding regions. If noncanonical G·U base pairs are at least partially compensatory for G·C interactions, then cytosine deamination mutation pressures may also account for the slight G+U bias noted in stem regions (Table 2). CpG methylation, prevalent in eukaryotes, is known to increase the rate of cytosine deamination (in this case directly to thymine, making recognition and repair of this mutation more difficult). This is consistent with the higher average G+U content observed for eukaryotic ssRNA compared to homologous proaryotic sequences (see Fig. 3 and compare the eukaryotic 23S, 16S, 5S rRNA, and tRNA G+U contents to their proaryotic homologues). These species-specific mutation pressures, when coupled with the effects of compensatory mutation described above, account well for the high variability in G+C content between species and the resulting G+C-directed axes parallel to Chargaff’s Axis.

Additionally, gene duplication events, though significantly increasing the length of ssRNA genes, will have little or no effect on composition. Bloch et al. (1983, 1985) have described possible duplication events relating tRNA and rRNA subsequences, suggesting that these two classes may have similar sequence (and hence base composition) characteristics due to common origins, despite their radical divergence in structure and function. In contrast, insertion, deletion, and recomb-
nutation of sizable elements can potentially alter the base composition of ssRNA coding regions drastically. These events may account for the disparity in composition between prokaryotic and eukaryotic sequences.

**Historical constraints**

The current understanding of mutational biases, however, cannot account for the low variability of G+A and G+U content nor the nearly universal bias in G+A. Although compensatory mutations occurring in stem regions allow for the large variability in G+C, it is not the case that mutations occurring in loop regions can always be compensated for by a change at a second site elsewhere in the molecule. The effects of mutation presumably act equally in all directions, but, without compensatory changes in loop regions, fewer mutations altering G+A and G+U content will be accepted, i.e., selection is more stringent in directions perpendicular to Chargaff's Axis. Congruent with our observations, this would prevent diversification in G+A and
TABLE 2. Relationship between nucleotide composition and secondary structure of various ssRNA.*

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Archaea</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>57.17</td>
<td>0.552</td>
<td>0.577</td>
<td>0.710</td>
<td>0.514</td>
<td>0.599</td>
<td>0.377</td>
<td>0.604</td>
<td>0.547</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>23S rRNA</td>
<td>3,019</td>
<td>44.39</td>
<td>11.46</td>
<td>1.32</td>
<td>0.00</td>
<td>57.17</td>
<td>0.552</td>
<td>0.577</td>
<td>0.710</td>
<td>0.514</td>
<td>0.599</td>
<td>0.377</td>
<td>0.604</td>
<td>0.547</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16S rRNA</td>
<td>1,476</td>
<td>50.41</td>
<td>8.40</td>
<td>1.90</td>
<td>0.81</td>
<td>61.52</td>
<td>0.560</td>
<td>0.556</td>
<td>0.524</td>
<td>0.683</td>
<td>0.521</td>
<td>0.567</td>
<td>0.357</td>
<td>0.613</td>
<td>0.452</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RNase P</td>
<td>315</td>
<td>56.51</td>
<td>3.81</td>
<td>0.00</td>
<td>0.00</td>
<td>60.32</td>
<td>0.470</td>
<td>0.559</td>
<td>0.495</td>
<td>0.593</td>
<td>0.503</td>
<td>0.529</td>
<td>0.288</td>
<td>0.648</td>
<td>0.448</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bacteria</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>56.60</td>
<td>0.542</td>
<td>0.586</td>
<td>0.506</td>
<td>0.667</td>
<td>0.523</td>
<td>0.558</td>
<td>0.378</td>
<td>0.670</td>
<td>0.437</td>
<td></td>
<td></td>
</tr>
<tr>
<td>23S rRNA</td>
<td>2,876</td>
<td>47.77</td>
<td>7.16</td>
<td>1.60</td>
<td>0.07</td>
<td>56.60</td>
<td>0.542</td>
<td>0.586</td>
<td>0.506</td>
<td>0.667</td>
<td>0.523</td>
<td>0.558</td>
<td>0.378</td>
<td>0.670</td>
<td>0.437</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16S rRNA</td>
<td>1,487</td>
<td>48.96</td>
<td>8.47</td>
<td>1.88</td>
<td>0.94</td>
<td>60.25</td>
<td>0.555</td>
<td>0.569</td>
<td>0.516</td>
<td>0.667</td>
<td>0.513</td>
<td>0.569</td>
<td>0.382</td>
<td>0.656</td>
<td>0.435</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5S rRNA</td>
<td>121</td>
<td>49.59</td>
<td>9.92</td>
<td>1.65</td>
<td>0.00</td>
<td>61.16</td>
<td>0.645</td>
<td>0.537</td>
<td>0.512</td>
<td>0.757</td>
<td>0.527</td>
<td>0.381</td>
<td>0.468</td>
<td>0.553</td>
<td>0.404</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RNase P</td>
<td>377</td>
<td>56.76</td>
<td>3.84</td>
<td>0.00</td>
<td>0.53</td>
<td>63.13</td>
<td>0.618</td>
<td>0.581</td>
<td>0.501</td>
<td>0.731</td>
<td>0.500</td>
<td>0.542</td>
<td>0.424</td>
<td>0.719</td>
<td>0.432</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Eucarya</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>59.79</td>
<td>0.479</td>
<td>0.550</td>
<td>0.540</td>
<td>0.566</td>
<td>0.525</td>
<td>0.609</td>
<td>0.339</td>
<td>0.589</td>
<td>0.430</td>
<td></td>
<td></td>
</tr>
<tr>
<td>23S rRNA</td>
<td>3,392</td>
<td>49.59</td>
<td>9.02</td>
<td>1.00</td>
<td>0.18</td>
<td>59.79</td>
<td>0.479</td>
<td>0.550</td>
<td>0.540</td>
<td>0.566</td>
<td>0.525</td>
<td>0.609</td>
<td>0.339</td>
<td>0.589</td>
<td>0.430</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16S rRNA</td>
<td>1,870</td>
<td>42.46</td>
<td>6.84</td>
<td>1.82</td>
<td>0.43</td>
<td>51.55</td>
<td>0.561</td>
<td>0.518</td>
<td>0.509</td>
<td>0.652</td>
<td>0.516</td>
<td>0.559</td>
<td>0.461</td>
<td>0.520</td>
<td>0.454</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>tRNA, PHE</td>
<td>76</td>
<td>57.89</td>
<td>5.26</td>
<td>7.89</td>
<td>2.63</td>
<td>73.67</td>
<td>0.539</td>
<td>0.539</td>
<td>0.526</td>
<td>0.642</td>
<td>0.566</td>
<td>0.547</td>
<td>0.304</td>
<td>0.478</td>
<td>0.478</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* N is the length of the sequence in nucleotides. W-C is the percent of base pairs that are A:U or C:G. GU, GA, and Other are the percent of other noncanonical base pairs. Total is the percent of bases in the entire sequence that are involved in base pair interactions. Base composition was calculated over the entire sequence (Full-length sequence), along paired bases (Stem) and unpaired (Loop) regions.

G+U content. Thus, the mean G+A and G+U values of a phylogenetic domain would be largely dependent on the values of the last common ancestor of that domain making G+A and G+U values at least partially constrained by historical contingency. Because G+C content evolves relatively rapidly, this measure is probably least sensitive to historical constraints. For example, the G+A and G+U content of Eucarya 16S rRNA are well constrained after billions of years of evolutionary history. The latest common ancestor of the Eucarya 16S rRNA probably had G+A and G+U values similar to the calculated mean values of contemporary sequences (0.524 and 0.527, respectively). Its not clear, however, where the latest common ancestor of extant life would have been located in the RNA simplex. Did other lineages from this time, which have since become extinct, also have G+A-rich compositions? Going back still further in evolutionary history to the time of the so-called RNA world (Gilbert, 1986; Joyce, 1989), did nascent RNA polymers have consistently biased compositions, or were they more randomly distributed?

FIGURE 7. Base composition biases among inferred base paired "stems" and unpaired "loops." Base composition (A: percent G+C content; B: G+A; C: G+U) among the total sequence (cross), inferred stems (circles), and inferred loops (square) of 10 ssRNA sequences representing five functional classes. The secondary structural elements of these ssRNA have distinct compositional biases (stems are G+C- and G+U-rich, loops are G+C-poor and G+A-rich), even though these sequences represent five different functional classes that share no sequence similarity. The mean G+C, G+A, and G+U content of the stem and loop regions of these 10 sequences were compared using the nonparametric Kolmogorov-Smirnov and Wilcoxon Signed-Rank tests. The null hypothesis, stating that the mean stem and loop compositions are equal, was rejected at the 99% confidence level for both G+C and G+U and at the 95% confidence level for G+A. Kolmogorov-Smirnov Test p-values are for G+C, 0.0001; for G+U, 0.0006; for G+A, 0.0169. Wilcoxon Signed-Rank Test p-values are for G+C, 0.0002; for G+U, 0.0004; for G+A, 0.0113. Although the proportion of bases involved in base pair interactions for these 10 sequences is remarkably constant among these varied sequences (averaging 60.52 ± 5.37%), the proportion of inferred base pairs decreases with sequence length. This trend may reflect actual changes in base pairing or may indicate that the number of base pair interactions in longer, more complex sequences is systematically underestimated.
Selection

Barring coincidence, neither mutational bias nor historical constraint can account for similar G+A and G+U values between distinct functional classes sharing no sequence similarity (and therefore no meaningful evolutionary history). If some or all of the 16 functional classes of ssRNAs studied here are, in fact, independently derived, then similar G+A and G+U values reflect evolutionary convergence (Doolittle, 1994) due to selection localizing extant populations of RNA sequences. For example, Bartel and Szostak (1993) isolated from $1.4 \times 10^{15}$ random sequences seven unique self-ligating ribozymes belonging to three classes (Table 1), each class representing sequences that apparently are evolved independently (Ekland et al., 1995). Despite their independent evolution from life on earth, the sequences characterized from each class are similarly G+rich, and are located among the natural distributions (Fig. 3F). This is a striking example of convergent evolution both within and between artificial and natural sequences that are known to share no evolutionary history. Although it would seem unlikely, nonselective compositional biases coincident with natural sequence biases could have been inadvertently imposed during the synthesis of the random library or in the amplification and mutation of the DNA templates. Controlling for these factors should be tractable experimentally in this and other artificial evolution systems (in this case, 165 of the 270 positions, or about two-thirds partial sequence data of 10 isolates of the initial pool show significant G+C biases (averaging $0.564 \pm 0.0349$, but unbiased G+A ($0.505 \pm 0.0326$) and G+U ($0.499 \pm 0.0402$) compositions (D. Bartel, pers. comm.). This enables the prospect of comparing directly, with suitable controls, a rapidly growing data set of artificial RNA sequence data to both naturally and other artificially evolved RNA lineages.

The inference that base composition is universally adaptive is corroborated by the biased distribution noted among secondary structural elements (Table 2; Fig. 7). Also, shorter sequences (roughly $N < 200$ nt) have smaller variance in G+A and G+U content than expected (Fig. 5), suggesting that selection culls from what would otherwise be a more variable distribution. However, this inference is contrary to current understanding in two respects. First, G+C content is almost always used alone in characterizing the base composition of RNA and, because of the large variability in G+C content, both within (Fleischmann et al., 1995) and between species, it has been assumed that overall base composition has little or no effect on the folding and function of ssRNA. Hence, Cantor and Schimmel (1980) state, “There is no evidence that the overall base composition of RNA or DNA correlates in any significant way with biological function.” However, it is not unreasonable to assume that the physicochemical differences among the four nucleotide bases impose differences on the biophysical properties of polymers having different base compositions. Although the sequential order of the bases is of obvious importance in higher-order structure formation, overall base composition might be expected to statistically bias attributes of the folding pathway. Thus, biophysical properties characterizing folded structure of RNA would, on average, differ from place to place within the RNA simplex. For example, the statistical distribution of secondary and tertiary structural elements (tetra loops, pseudoknots, coaxial helices, parallel pairs, unilaterally and bilaterally bulged residues) might be expected to change as a function of the percent G residues (Gutell et al., 1994). Currently, we are using computer simulations to systematically map within the RNA simplex the anisotropic distribution of several thermodynamic variables characterizing folded RNA structure.

Second, similar composition biases among unrelated sequences implies that these biases are largely independent of functional and phylogenetic constraints. By analogy with protein folding (Frauenfelder & Wolynes, 1994; Bryngelson et al., 1995; Wolynes et al., 1995; Li et al., 1996), a prerequisite property for almost any molecular function is the expedient formation of unique and stable, folded conformations (Herschlag, 1995; Uhlenbeck, 1995). We hypothesize that similar G+A and G+U biases may reflect generic requirements for stable, unique, and functional conformations. Because purines are noted for their exceptional stacking interactions (Saenger, 1984), G+A-rich loop regions noted in Table 2 and Figure 7 could contribute to conformational stability despite the lack of Watson–Crick pairing. For example, Cate et al. (1996), found a preponderance of purines among long-range, stacking interactions in the crystal structure of the P4–P6 domain of the Tetrahymena thermophila self-splicing intron. The utility of such biases would be analogous to the role of G+C biases of stems that increase thermostability in RNA from hyperthermophilic organisms (Brown et al., 1993). A specific province of composition space, representing a balance between G+C, G+A, and G+U contents, would contain a higher “density” of well-folding sequences. This province would therefore act as an attractor as genealogically independent ssRNA and artificial ribozymes originated and diversified. For example, compositional biases could have favored an increase in the structural stability of random and poorly adapted ssRNA sequences condensing in hostile, nascent environments. Thus, the province of composition space occupied by contemporary sequences could have preferentially supported the independent origin of functionally distinct ssRNA. Originating in the same compositional province, ssRNA unrelated by sequence similarity and ancestry would nonetheless have similar compositional biases. However, this does not preclude the possibility of the origin of functional sequences in other provinces of composition space, followed by
subsequent evolutionary convergence to the contemporary composition values. Evolutionary diversification via compensatory mutations would have occurred in directions parallel to, but displaced from Chargaff’s Axis, yielding the observed, axis-like, distributions.

U1-U5 snRNAs are the only functional classes observed to have mean G+A compositions smaller than 0.5. These anomalous G+A values presumably reflect some highly specialized adaptation within the spliceosome. For example, the low G+A values may be indicative of a sparsity of long-range, purine-rich contacts within these snRNAs. However, the most phylogenetically conserved snRNA moiety, U6 snRNA (Madhani & Guthrie, 1992; McPheeters & Abelson, 1992; Wise, 1993), has G+A and G+U values that are typical of group I, group II, and P RNA ribozymes. Because the U6 snRNA is thought to constitute the catalytic core of the spliceosome (along with the U2 snRNA), it may be the case that the purine-rich composition of the U6 snRNA sequence is necessary for the two-step splicing reaction (analogous to self-splicing group II introns) carried out by the spliceosome. The base composition of U6 snRNA would therefore have been conserved, whereas the other snRNA sequences were allowed to adapt specialized, G+A-poor compositions.

**Conclusion**

Similarities that arise from common ancestry can be controlled for by comparing independent lineages of molecular sequences. Comparative statistical analyses of this sort will begin to define characteristics of ssRNAs that are common to any well-evolved, functional RNA sequence in contrast to conventional comparative sequence analyses that define characteristics unique to individual lineages. ssRNA sequences derived from artificial evolution experiments are genealogically independent of life on earth, and can therefore act as replicate experiments for the origin of functional RNA in that properties common to both natural and artificial sequences must be due to convergence and not historical constraint. Only in this case can the confounding influence of genealogy be completely eliminated from evolutionary inference. In this study, we have documented similar base composition distributions among ssRNA sharing no evolutionary history. From this result, we infer that G+A and possibly G+U biases are functionally and phylogenetically universal adaptations. Comparative statistical analyses, however, need not be restricted to base composition. Secondary (Fontana et al., 1993) and tertiary structures, as well as experimentally obtained biophysical and functional properties (Kuo & Cech, 1996), are all amiable to comparison. As more sequence data becomes available, comparative statistical analyses between sequences sharing little if any evolutionary history will become increasingly germane to the general understanding of molecular structure, function, and evolution.

**MATERIALS AND METHODS**

**The RNA simplex**

The number of possible RNA sequences of length $N$ is given by $4^N$ (this is the size of the so-called sequence space; Smith, 1970; Hamming, 1980; Eigen, 1992). Each of these sequences can be classified into a compositional class denoted by its composition vector. The space of all possible RNA composition vectors is constrained to the volume of tetrahedron. This is a three-dimensional projection of high-dimensional sequence space in that all $4^N$ possible sequences are projected onto

$$C = \binom{N + 3}{N}$$

compositional classes. For example, for $N = 20$, all $1.1 \times 10^{12}$ sequences are partitioned among $C = 1,771$ compositional classes. The “density” of a composition class $c$ (the number of sequences belonging to the class $c$) is given by the multinomial distribution:

$$p_c = N!/(A! \times C! \times G! \times U!),$$

where $A$, $C$, $G$, and $U$ specify the number of each residue within a sequence of $N$ total residues; $N = A + C + G + U$. $p_c$ can be summarized by calculating the Shannon entropy (Shannon, 1948) of a composition vector $c$ as:

$$H_c = -(A/N \log_2 A/N) + (C/N \log_2 C/N) + (G/N \log_2 G/N) + (U/N \log_2 U/N).$$

Note that the density of sequences increases enormously toward the center of the simplex (from 1 at the homopolymer composition vectors to more than $10^{10}$ at the isoheteropolymer composition vectors, when $N = 20$). It is for this reason that a random walk in sequence space, where each sequence is equally likely to be visited, will be confined near the high entropy, isoheteropolymers.

Although relatively new to ssRNA, there is, in fact, precedence for comparative composition analysis in the protein literature. Most relevant to our study, Chou (1995), using a representative compilation of distantly related protein sequences, has shown that the amino acid composition is an excellent predictor (95.3% accuracy) of protein structural class. Intriguingly, the interpretation is that the different structural classes of proteins are localized in distinct provinces of protein composition space [i.e., the (20-1)-dimensional protein simplex]. Eisenhaber et al. (1996a, 1996b) have critically reviewed Chou’s work, but nonetheless conclude that “secondary structural content of a protein is determined mainly by the amino acid composition.”

**Visualization software**

The unit simplex was visualized with Mage 4.4, an interactive molecular graphics software package. Composition
data are visualized by specifying the calculated composition vectors as the fraction of A, C, and G (the fraction U being implicit) in the Kinemage data file. The latest version of Mage and the Kinemage data files used in this work can be obtained via anonymous ftp, at: http://www.santafe.edu/~pth/simplex.html.

Database construction

Composition vectors and base composition statistics of ssRNA sequences (Table 1) were calculated using a combination of standard Unix based utilities and Microsoft Excel 4.0. Composition vectors were calculated from full-length sequence data compiled from various Internet sources: 23S and 16S rRNA sequences were obtained from http://rrna.uia.ac.be; 5S rRNA sequences were obtained from http://cammsg3.caos.kun.nl; snRNA were obtained from http://pegasus.uhct.edu/uRNADB/uRNADB.html; tRNA sequences were obtained from ftp://ftp.ebi.ac.uk/pub/databases/trna; P RNA sequences obtained courtesy of Jim Brown; 18S rRNA sequences obtained courtesy of Kevin Peterson; Group I, Group II, and hammerhead ribozymes obtained through GenBank searches; additional Group I and Group II sequences obtained from Schmelzer and Schwyen (1986), Schmidt et al. (1990), and Green and Szostak (1992). Only sequences greater than 90% complete are included. In an attempt to produce phylogenetically representative samples of molecular diversity, each genus is represented by a single, randomly chosen species (except in the tRNA data set, where all available sequences are plotted).

Base pairing data for Table 2 were obtained by inspection, using inferred secondary structures (putative tertiary base pair interactions were also counted) from the literature (Roberts et al., 1974; Noller, 1984; Gutell et al., 1993, 1994) and the Ribonuclease P Database, available at ftp://iubio.bio.indiana.edu/molbio/rnase-p/home.html. The sequences (plotted from left to right in Fig. 7) are: (Archaea) 235 rRNA, Methanobacterium thermoautotrophicum, X16364; 16S rRNA, Methanobacterium fomicicum, M36508; P RNA, Sulfofobus acidoalcaldarius, L13597; (Bacteria) 23S RNA, Synecococcus sp. 6301, X00512; 16S RNA, Synecococcus sp. 6301, X03538; 5S rRNA, Escherichia coli, M24300; P RNA, E. coli, X00211; (Eucarya) 23S rRNA, Saccharomyces cerevisiae, J01355; 16S rRNA, Homo sapiens, K03432; tRNA, yeast, M10263.

Random “RNA” sequences were generated computationally by choosing “bases” from a uniform distribution (each base occurs with a frequency of 0.25). These random sequence files were then treated like biological sequences in order to calculate their base composition statistics. Complete data sets with accession numbers are available via anonymous ftp, at: http://www.santafe.edu/~pth/simplex.html.

ACKNOWLEDGMENTS

We thank Francois Michel for directing us to Group II data, Kevin J. Peterson for 18S RNA data and Jim Brown for P RNA data sets. We thank D. Richardson and J. Richardson for computing resources; A. Andrews, T. Tollesbøl, M. Word, D. Kenan, J. Keene, M. Geysen, J.W. Schopf, and an anonymous reviewer for helpful discussions and advice. This work was supported by the Center for the Study of the Evolution and Origin of Life, Diversity Biotechnology Consortium, the Combinatorial Sciences Center at Duke University, the Santa Fe Institute and NASA grant NAGW2147.

Received December 5, 1996; returned for revision January 14, 1997; revised manuscript received May 1, 1997

REFERENCES


