

# A Parameterization of RNA Sequence Space

**ERIK SCHULTES**

*Department of Earth and Space Sciences, University of California at Los Angeles, Los Angeles, CA 90024, and Combinatorial Sciences Center, Duke University Medical Center, Durham, NC 27710*

**PETER T. HRABER**

*Department of Biology, University of New Mexico, Albuquerque, NM 87131, and National Center for Genome Resources, 1800 Old Pecos Trail, Suite A, Santa Fe, NM 87501*

**THOMAS H. LABEAN**

*Combinatorial Sciences Center, Duke University Medical Center, Durham, NC 27710, and Department of Biochemistry, Duke University Medical Center, Durham, NC 27710*

*Received June 23, 1998; accepted November 6, 1998*

*RNA polymers are constructed from four distinct nucleotide bases. The sequence of these nucleotide bases determines both the folded conformation and the biological function of RNA. It recently has been established that disparately related functional classes of evolved RNA possess similar base composition biases despite a lack of sequence similarity, folded structure, or metabolic function. We have proposed that intrinsic constraints on RNA structure have imposed this convergent evolution in base composition. Here, we test this hypothesis by first calculating the distribution of the mean thermodynamic stability of random RNA sequences as a function of base composition. Then, using a model describing mutation (as a random walk in sequence space) and selection (which tends to increase thermodynamic stability), we relate the computed underlying distribution of conformational stability to empirically derived, tRNA and 5S rRNA sequence data. We find a close correspondence between predicted and observed distributions of base composition. © 1999 John Wiley & Sons, Inc.*

**Key Words:** sequence space, RNA evolution, RNA simplex, evolutionary trajectories, evolutionary attractor, intrinsic evolutionary constraints

**K**auffman has long argued that “biologists have, as yet, no conceptual framework in which to study an evolutionary process that commingles both self-organization and selection” [1, quoted from 2]. Here, we present a candidate framework relating self-organization and selection in the evolution of single-stranded ribonucleic

acids (RNAs). Not only is RNA of significant biological interest (having both the ability to store genetic information and perform specific metabolic activities), but modern methods in molecular and structural biology make RNA the ideal model system in which to address both theoretically and experimentally the relationship between self-

organization and selection [3]. Our approach is motivated by a recent comparative analysis of 15 distinct functional classes of RNA, which documented similar nucleotide base composition biases among disparately related RNAs sharing little or no sequence similarity [4]. We develop the framework in three sections. First, we define a parameterization scheme for RNA sequence space based on nucleotide composition. Second, using this parameterization scheme, we compute among random sequences the statistical distribution of thermodynamic stability (i.e., self-organization) of RNA secondary structures as a function of base composition. Third, we perform simulations of RNA evolution incorporating the calculated distribution of self-organization, leading to predictions of optimal base composition values for adapted molecular sequences. These results are compared with empirically derived transfer RNA and 5S ribosomal RNA sequence data.

### PARTITIONING RNA SEQUENCE SPACE

The framework relating selection and self-organization is based on the notion of a molecular sequence space, first proposed in the context of proteins [5–7], and subsequently extended to RNA [8–17]. In this high-dimensional space, each point uniquely represents a possible RNA sequence, and neighboring points represent sequences that differ by single base substitutions. Because the biochemical functions of RNA sequences are dependent on folded conformation, the ideal ordering scheme for RNA sequence space would partition the space in such a way that sequences from the same partition would support similar conformational properties. For example, sequences from the same partition might have folded conformations of similar thermodynamic stability. Inspired by Langton's  $\rho$ -space, which was used to parameterize cellular automata rule space [18,19], we impose a parameterization scheme on the space of RNA sequences using nucleotide base composition. The base composition of an arbitrary RNA sequence having  $N$  nucleotide residues can be denoted by its normalized composition vector  $(A/N, C/N, G/N, U/N)$ , where  $A, C, G,$  and  $U$  are the number of occurrences of each base in the sequence. All possible partitions can be geometrically represented as points within the volume of a tetrahedron—the so-called RNA simplex [4]. The four homopolymers (poly-A, poly-C, poly-G, poly-U) reside at the vertices, while sequences having an equal number of the four nucleotides reside at the center of gravity of the tetrahedron, equidistant from the homopolymers. These heteropolymers, having the composition vector  $(0.25, 0.25, 0.25, 0.25)$ , are referred to as isoheteropolymers. We have visualized the RNA simplex (Figure 1) using an interactive graphics package [20,21]. A more convenient denotation of the composition vector is the specification of the fraction  $G+C, G+A,$  and  $G+U$  contents of RNA sequences. These measures define three mutually perpendicular compositional “gradients” within the

FIGURE 1

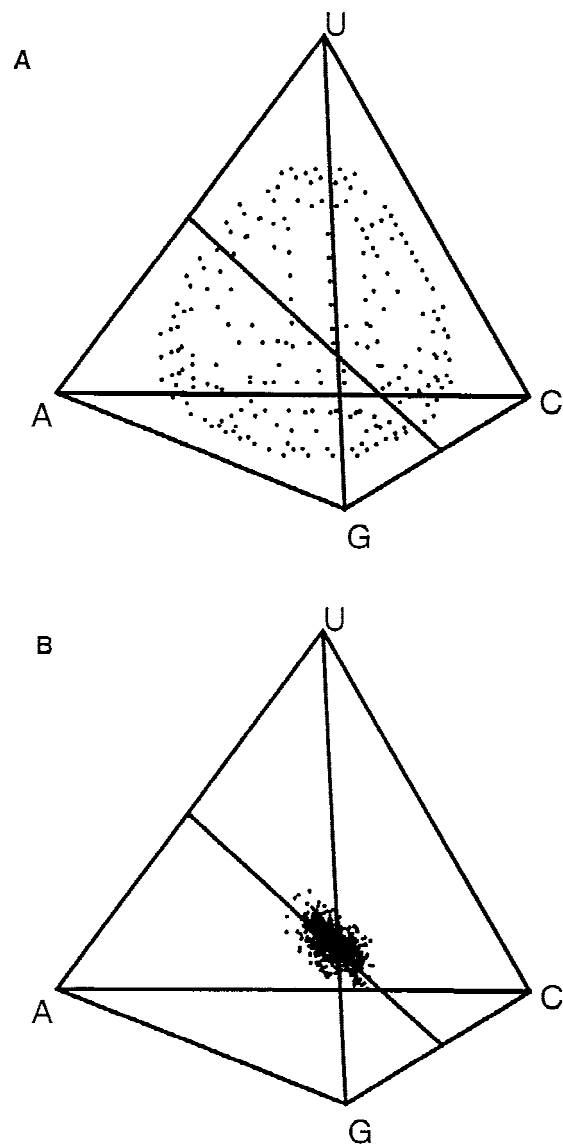


FIGURE 1. RNA sequence space can be partitioned by placing in the same partition all sequences having the same nucleotide base composition. The density of a partition (the number of sequences that belong to that partition) is described by the Shannon entropy,  $S$ , calculated from the composition vector. (A) Depicted are all composition vectors where  $1.3 < S < 1.5$ . Composition vectors of equal entropy form concentric “spheres” centered on the isoheteropolymers. For moderately long sequences, the density increases drastically toward the high-entropy isoheteropolymers. Also depicted is Chargaff's Axis, the locus of composition vectors where the number of Watson-Crick bases are equal. Chargaff's Axis extends from the midpoints of the AU and CG edges through the isoheteropolymers. (B) Plotted is the distribution of 928 cytoplasmic tRNA molecules from organisms spanning the three domains of the universal tree of life [4]. The mean base composition of these 928 tRNA sequences is given in the text.

simplex and, taken together, can uniquely specify position within composition space. The RNA simplex has been described in detail elsewhere [4]. Here, we summarize some important points.

### The Importance of Base Composition

Though base composition statistics provided key information in the elucidation of the double-helix structure of genomic DNA [22], recent advances allowing efficient DNA sequencing have shifted attention away from base composition statistics [23]. Since RNA folding is known to be sequence specific, base composition may seem an overly crude parameter in characterizing RNA conformation. However, this parameterization scheme has two justifications. First, the four nucleotide bases are sterically and chemically distinct: Thus, different ratios of the four bases would be expected to statistically influence the folding properties of RNA. Second, despite the well-documented variability in G+C content, we have demonstrated among 2,800 naturally evolved sequences a significant localization of base composition in a G+A-rich and G+U-constricted province of the simplex (see [4]; as an example here, see the tRNA and 5S rRNA distributions in Figures 1B, 5B, and 5D). The great disparity in sequence and structure among 15 functional classes of RNA investigated imply that these molecules share little, if any, evolutionary history. The coincidence of base composition among these sequences, therefore, suggests adaptive convergence, from which we infer that base composition plays an important role in the evolution of RNA structure and function.

### The RNA Simplex: Projection of Sequence Space

The RNA simplex is more than a concise graphical representation of base composition. Because all possible sequences are partitioned into the possible composition vectors, the simplex can be considered a low-dimensional projection of an otherwise high-dimensional sequence space. As such, the simplex is a rather complicated object. For example, molecules having similar sequences will be located near each other in the simplex. However, two sequences differing at every position will be maximally distant in sequence space but could nevertheless share the same composition vector. This is, in fact, the case in observed distributions, where, for example, tRNA and 5S rRNA base compositions demonstrate significant overlap despite a lack of sequence similarity.

The density of composition vectors (number of sequences belonging to a particular composition vector) changes enormously from the outer edges of the simplex toward the isoheteropolymers. For example, each of the four homopolymer composition vectors contain only a single sequence, whereas even for relatively short RNA sequences having only 100 nucleotides, over  $10^{57}$  sequences (of the  $1.6 \times 10^{60}$  possible sequences) belong to the isohet-

eropolymer composition vector. The distribution of density across the RNA simplex is effectively summarized by the Shannon information entropy [24], calculated for each composition vector,  $c$ :

$$S_c = -[(A/N \log_2 A/N) + (C/N \log_2 C/N) + (G/N \log_2 G/N) + (U/N \log_2 U/N)] \quad (1)$$

RNA sequences map trajectories in sequence space as they undergo evolutionary modification by mutation and selection. Trajectories involving changes in sequence composition can be observed in the RNA simplex. If the forces of mutation and selection resulted in an evolutionary modification identical to a random walk in sequence space (i.e., such that each sequence was equally likely), then the observed base composition of evolved sequences would eventually converge toward the isoheteropolymers simply because this partition contains more sequences than any other partition. Since the base composition of evolved RNA does not conform to the high entropy case, a random walk is at best an incomplete model for the evolution of RNA. Elsewhere, we have proposed that the trajectories of evolving molecules are influenced by a nonuniform distribution of folding capacity with respect to base composition [4]. In order to test this hypothesis, we use the RNA simplex parameterization scheme and computational approximations of RNA folding to map the distribution of mean thermodynamic stability across sequence space.

## MAPPING SELF-ORGANIZATION IN RNA SIMPLEX

### Conformational Order and Self-Organization in RNA

**D**riven by the release of free energy in the formation of intramolecular contacts, RNA polymers under appropriate conditions, either evolved or random, spontaneously fold into complicated three-dimensional structures [25–29]. The degree of order (or disorder) in the folded structure can be measured by several methods [3,10,15,30–33]. The simplest quantification of conformational order is to compute the base-pairing propensity,  $P$ , defined as the number of hydrogen bonded base-pair interactions in a folded RNA, normalized to the length of the sequence in nucleotides,  $N$ . The most prevalent interactions (referred to as Watson-Crick base-pairs) occur between adenosine (A) and uracil (U) bases, and cytosine (C) and guanine (G) bases.  $P$  ranges from 0.0 (no base-pairing) to 0.5 (maximum possible base-pairing). Random sequences display a range of base-pairing propensity, some having well-ordered conformations, others having poorly ordered conformations. The mean base-pairing propensity,  $\langle P \rangle$ , of an ensemble of random sequences is a measure of the degree of spontaneous order, or self-organization among that ensemble [3]. Thus, self-organization is a statistical property of the ensemble of random sequences; individual sequences may be more or less ordered than the average but will also be more

rare. The degree to which a folded conformation is ordered depends precisely on the sequential order of its nucleotide bases. However, it is reasonable to expect the degree of self-organization to be statistically dependent on base composition.

### Estimating Self-Organization: Mean-Field Approximation

The simplest analytical approximation to RNA secondary structure formation averages the mean Watson-Crick base-pairing propensity over long, random RNA sequences, ignoring the effects of the linear ordering of the sequence. As a further simplification, A·U and C·G base-pairs are treated equally and noncanonical base-pairing is disallowed. This mean-field approximation estimates the mean base-pairing propensity of random sequences as the sum of the independent probabilities that an A will pair with a U and that a C will pair with a G elsewhere in the molecule. These pair-wise probabilities are proportional to the frequency with which each base occurs in the sequence (e.g., increasing the number of G residues would be expected to increase the number of C·G base-pairs) and can be derived directly from the normalized composition vector as:

$$\langle P_c \rangle = 2 \cdot ((A/N \cdot U/N) + (C/N \cdot G/N)) \quad (2)$$

This formulation (referred to as the “stickiness” of the sequence [10]) allows  $\langle P \rangle$  to be calculated for each composition vector in the RNA simplex.  $\langle P_{min} \rangle = 0.0$  occurs for composition vectors where no Watson-Crick complements coexist (e.g., at the vertices or along edges other than the AU and CG edge).  $\langle P_{max} \rangle = 0.5$  occurs for only two composition vectors: the midpoint of the AU edge (0.5, 0.0, 0.0, 0.5) and the midpoint of the CG edge (0.0, 0.5, 0.5, 0.0). The locus of points joining these midpoints extends through the isoheteropolymers and represents the RNA equivalent of Chargaff's Rule (A = U and, simultaneously, C = G). We refer to this locus as Chargaff's Axis. Though a maximum of 100 percent base-pairing is allowed along the length Chargaff's Axis,  $\langle P \rangle$  decreases from  $\langle P \rangle = 0.5$  at the endpoints of Chargaff's Axis to  $\langle P \rangle = 0.25$  at the isoheteropolymers. This decrease in  $\langle P \rangle$  from the endpoints of Chargaff's Axis to the isoheteropolymers can be interpreted as the increasing preponderance of A and U nucleotides “frustrating” the base-pairing of C and G nucleotides and vice versa [34–37]. Consequently, we quantify the degree of intramolecular frustration,  $\langle F_p \rangle$ , in the mean-field model as a simple rescaling of  $\langle P \rangle$ , such that:

$$\langle F_{pc} \rangle = 2 \cdot (0.5 - P_c) \quad (3)$$

Hence, self-organization can also be defined as the spontaneous minimization of mean frustration from a maximum

value of 1.0. This rescaling is convenient for comparing the mean-field model to the thermodynamic approximation in the next section.

Figure 2A represents the distribution of mean base-pairing propensity over 1,771 composition vectors spanning the RNA simplex. Each composition vector shown represents a 5 percent change in base composition. Surfaces of equal  $\langle P \rangle$  values form stability fields having a twofold symmetry about Chargaff's Axis. Intermediate stability fields have an hourglass shape, describing a decrease in mean base-pairing propensity due to frustration from the endpoints of Chargaff's Axis toward the isoheteropolymers. This calculation provides first-order estimates of the spontaneous formation of base-pair interactions among random RNA heteropolymers as a function of base composition.

### Estimating Self-Organization: Thermodynamic Approximation

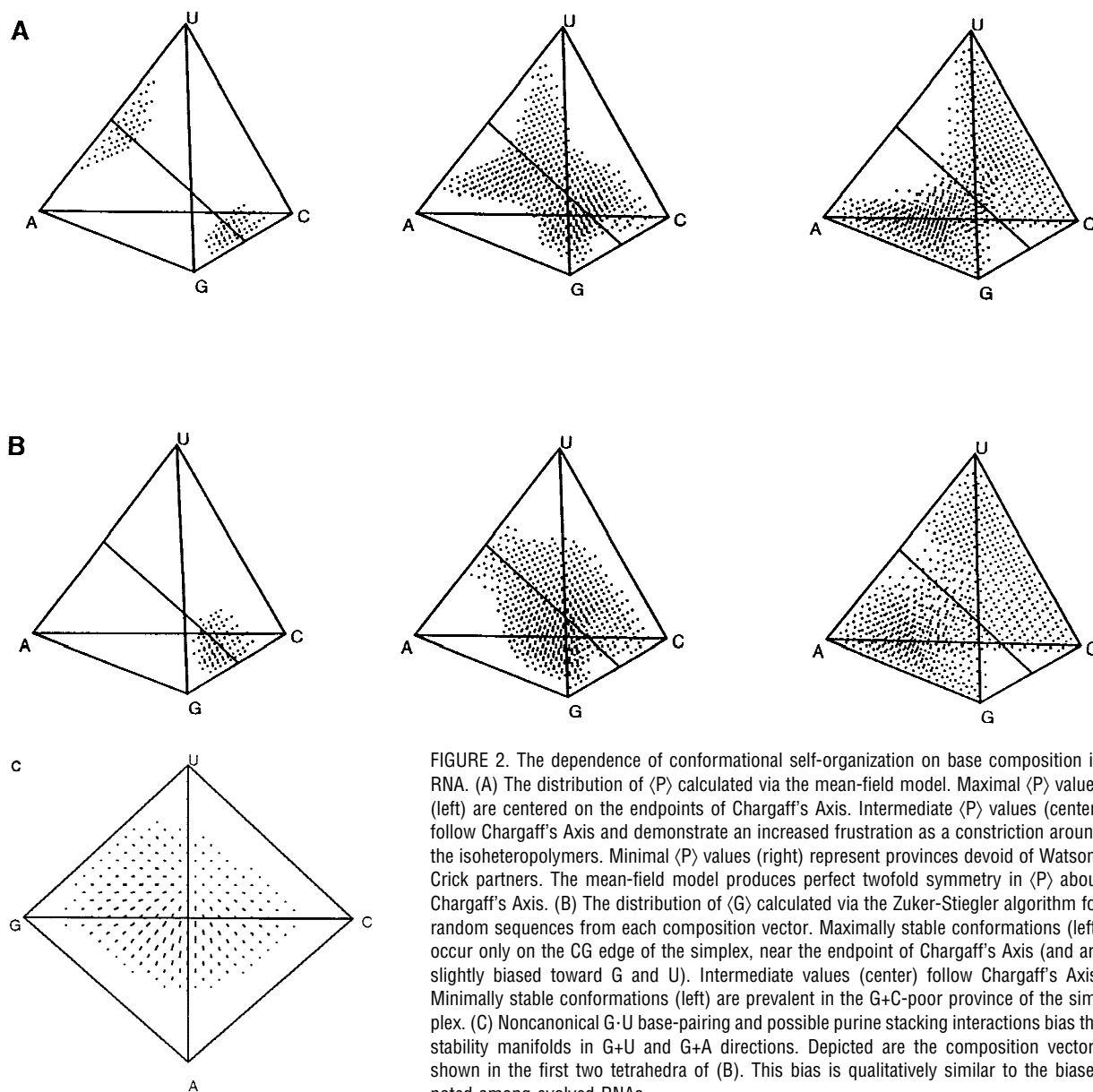
RNA secondary structure accounts for the majority of the free energy of structure formation and plays an important role in the interpretation of RNA function and evolution [14]. A more realistic approximation to RNA folding can be derived from folding algorithms designed to map nucleotide sequence to minimal free-energy secondary structures using empirically derived thermodynamic parameters. Here, we employ a standard, efficient algorithm used in similar analyses [12,38–41] for calculating RNA secondary structures from the nucleotide sequence to more accurately estimate the distribution of self-organization of RNA across sequence space.

One hundred random sequences of length  $N=100$  were generated for each of 1,771 composition vectors spanning the simplex. The minimum free-energy structure for each sequence was calculated at 37°C. The free energy ( $\Delta G$ , a measure of the stability of the folded conformation in Kcal/mol) of each of the 100 sequences was then averaged, providing an estimate of the mean thermodynamic stability of random sequences from each composition vector,  $\langle \Delta G_c \rangle$ . In order to compare these calculations more directly with the mean-field model, we normalized these free-energy values by dividing them by the free energy of what we believe to be the lowest energy conformation possible: a hairpin structure given by the sequence of alternating GC bases,  $(GC)_{50}$  having  $\Delta G_{(GC)_{50}} = -137.47$  Kcal/mol. The mean frustration,  $\langle F_{\Delta G} \rangle$ , was calculated as the difference of this quotient from unity:

$$\langle F_{\Delta G} \rangle = 1 - \left( \frac{\langle \Delta G_c \rangle}{\Delta G_{(GC)_{50}}} \right) \quad (4)$$

Perhaps surprisingly, this simulation method, which incorporates the details of sequential ordering, yields distributions that are qualitatively similar to that of the much simpler mean-field model (Figure 2B). However, unlike the

**FIGURE 2**



**FIGURE 2.** The dependence of conformational self-organization on base composition in RNA. (A) The distribution of  $\langle P \rangle$  calculated via the mean-field model. Maximal  $\langle P \rangle$  values (left) are centered on the endpoints of Chargaff's Axis. Intermediate  $\langle P \rangle$  values (center) follow Chargaff's Axis and demonstrate an increased frustration as a constriction around the isoheteropolymers. Minimal  $\langle P \rangle$  values (right) represent provinces devoid of Watson-Crick partners. The mean-field model produces perfect twofold symmetry in  $\langle P \rangle$  about Chargaff's Axis. (B) The distribution of  $\langle G \rangle$  calculated via the Zuker-Stiegler algorithm for random sequences from each composition vector. Maximally stable conformations (left) occur only on the CG edge of the simplex, near the endpoint of Chargaff's Axis (and are slightly biased toward G and U). Intermediate values (center) follow Chargaff's Axis. Minimally stable conformations (right) are prevalent in the G+C-poor province of the simplex. (C) Noncanonical G·U base-pairing and possible purine stacking interactions bias the stability manifolds in G+U and G+A directions. Depicted are the composition vectors shown in the first two tetrahedra of (B). This bias is qualitatively similar to the biases noted among evolved RNAs.

mean-field model, the thermodynamic approximation incorporates the realistic differences in the thermodynamic properties of the four bases. For example, the number of hydrogen bonds between A·U base-pairs (2) is less than C·G base-pairs (3). Thus, C·G bonds contribute more to the stability of the folded structure than do A·U bonds. Also, the folding algorithm used in this study incorporates the non-canonical G·U base-pair interaction. The physiochemical asymmetry of the four bases imposes an asymmetry across the stability fields in the simplex. These asymmetric interactions result in four important deviations from the mean-

field model. First, the relatively weaker A·U interactions succumb to C·G interactions, resulting in a globally optimal stability, located on the CG edge. The bottleneck character of the intermediate stability fields are correspondingly deformed with obvious biases of increased stability toward the G+C-rich compositions. Second, there is a slight bias of increased stability in G+A-rich (i.e., purine-rich) compositions. These biases are likely to be the result of highly stable purine-purine stacking interactions [42] found in abundance in sequences from purine-rich provinces of the simplex. Third, there is a slight bias of increased stability in

sequences having G+U-rich compositions (Figure 2C). This stability is presumably the result of increased base-pairing due to the noncanonical G·U interactions. The resultant effect of these intrinsic factors shifts the global minimum frustration from (0.0, 0.5, 0.5, 0.0) as in the mean-field case to a neighboring composition vector, still on the CG edge but slightly enriched in G (0.0, 0.45, 0.55, 0.0). Lastly, both the mean-field and thermodynamic approximations demonstrate that, because of Watson-Crick base-pairing, the vicinity of Chargaff's Axis is the most well-ordered province of the RNA simplex. Those composition vectors lacking in Watson-Crick partners (e.g., near the AG edge) are disordered, forming only a minimal number of base-pair interactions and having minimally stable conformations.

Hence, the directions perpendicular to Chargaff's Axis, extending toward the AG, AC, CU, and GU edges of the tetrahedron, represent the most direct routes from statistically ordered to disordered regions of RNA sequence space. Composition vectors lying between Chargaff's Axis and these four edges define four order-disorder manifolds. On these manifolds, base composition acts as a course-grained control parameter for intramolecular hydrogen bond interactions between the nucleotide bases [9]. Figure 3 depicts the order-disorder manifold extending from Chargaff's Axis to the AG (purine) edge of the tetrahedron. The difference between the order-disorder transitions between the G+C-rich and the G+C-poor compositional regimes of this manifold occur because the A·U base-pair is a weaker interaction than the C·G base-pair. The sharpest order-disorder transition (determined by the slope of the inflection point in the sigmoidal transition curve) occurs in the G+C-rich province of the purine-manifold, corresponding to the same octant in which biological sequences are most frequently observed. However, taken by itself, the geometry of the calculated stability fields does not describe the distributions of evolved sequences. In the next section, we propose a model of molecular evolution where the underlying distribution of spontaneous organization in RNA contributes to the observed localization of evolved sequence data.

### **PREDICTING EVOLUTIONARY TRAJECTORIES: SELECTION, MUTATION, AND EVOLUTIONARY POTENTIAL**

Each composition vector has an intrinsic potential to yield well-adapted sequences during an evolutionary search. The *evolutionary potential* of a composition vector will depend on the probability that such a vector will be sampled by mutation and the probability that it will yield useful sequences. These expectations are estimated later for each of 1,771 composition vectors spanning the RNA simplex, producing a theoretical landscape that would direct the flow of evolving populations in the simplex. "Basins" of evolutionary potential act as attractors, localizing the base composition of adapted sequences even though their primary structure continues to evolve.

### **Mutation: Probability of Sampling a Composition Vector**

The RNA genes encoded by the genomic DNA are subject to mutations in their storage and replication. Assuming random mutation, uniform over the four nucleotide bases, sequence composition will converge to the isoheteropolymers ( $\langle G+C \rangle = \langle G+A \rangle = \langle G+U \rangle = 0.5$ ) with a standard deviation that decreases as the square root of sequence length [4]. Since base-substitution is known to be nonuniform in several senses [43–46], assumptions of ergodic mutation would appear untenable. Yet the double-stranded nature of the genomic DNA constrains any nonuniformity in the mutation rate of the four nucleotides as changes in only G+C content (i.e., as changes parallel to Chargaff's Axis) [4]. For organisms with double-stranded DNA, average G+A and G+U contents cannot be altered from 0.5 regardless of the nonuniformity of the mutation rates. This constraint is due to the fact that mutations occur equally likely on both strands of the DNA. During DNA replication, the complementary strand will accommodate mutations with the appropriate Watson-Crick base-pair. Hence, each strand of the DNA genome accumulates on average an equal number of A and T mutations and C and G mutations. This phenomenon has been directly observed in organisms for which complete genomic data exist (e.g., [47,48]). Thus, at least for G+A and G+U contents, an assumption that mutation tends to uniformly disorder RNA sequences is a reasonable first approximation to the multiple factors influencing mutation bias.

### **Selection: Probability That a Composition Vector Contains Functional Sequences**

In general, RNA folding is frustrated by numerous conflicting interactions that cannot be simultaneously satisfied [34,37]. In the context of protein folding, Frauenfelder and Wolynes have proposed a principle of minimum frustration (PMF) stating that evolved sequences have been selected such that the number of frustrated interactions have been minimized compared to that of random heteropolymers [35,49]. In the case of RNA, the PMF has been verified for nine disparately related functional classes [3,31,32]. In these studies, the free energies of evolved RNA sequences (taken here as a proxy for frustration) have been shown to be significantly lower than the free energies of random sequences. This result is relevant because locations in the RNA simplex characterized by spontaneously low frustration values will preferentially support the origin and evolutionary diversification of RNAs. Given the underlying distribution of frustration across the RNA simplex, it follows from this assumption that selection would tend to drag sequence composition toward the endpoint(s) of Chargaff's Axis where base-pairing propensity is maximized. This assumption appears absurd in the extreme case where se-

**FIGURE 3**

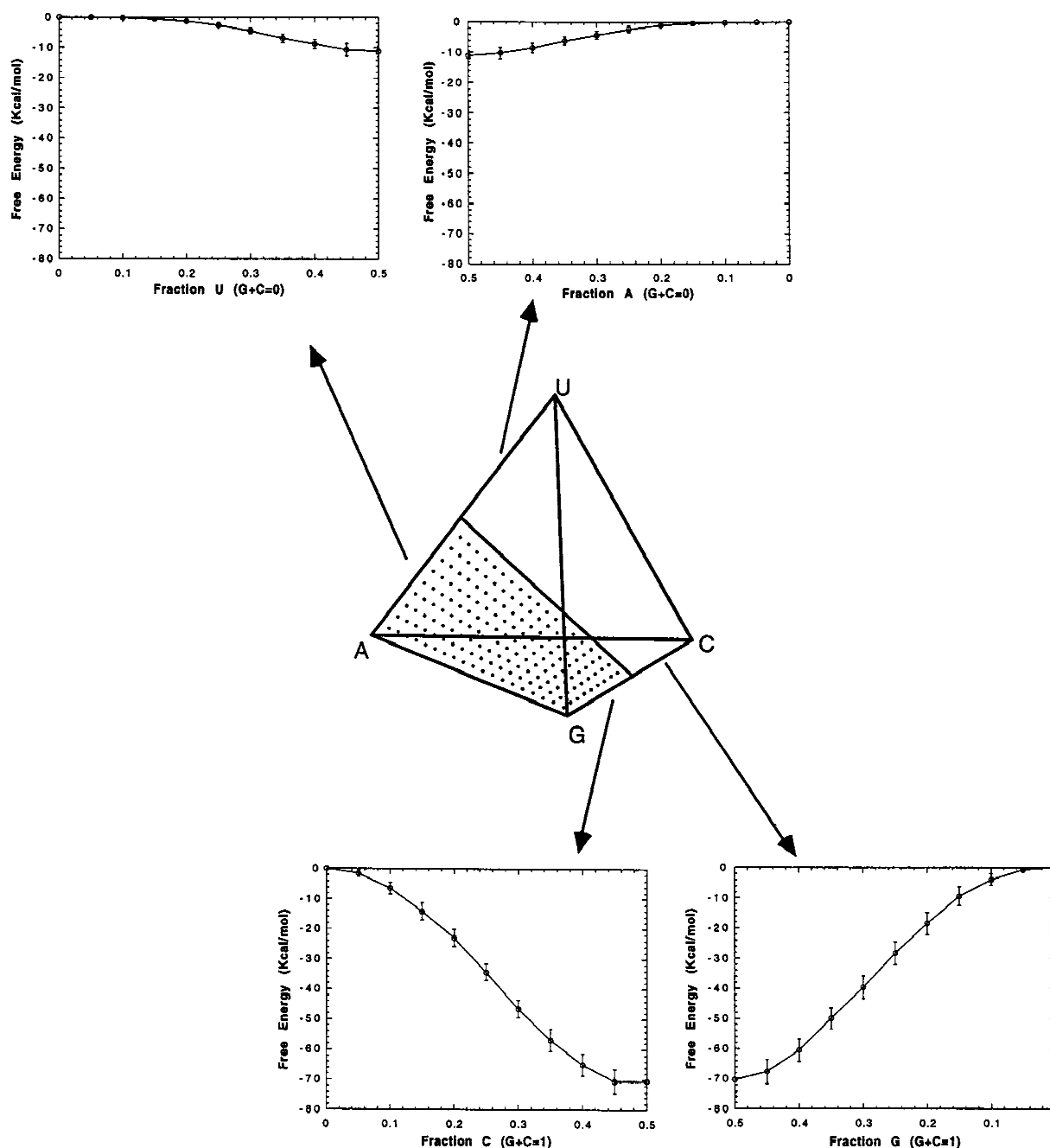


FIGURE 3. Depicted within the simplex is the purine-manifold, extending from Chargaff's Axis toward the AG edge. To demonstrate these order-disorder transitions quantitatively, consider a random sequence located at the C=G=0.5 endpoint of Chargaff's Axis. This sequence may be "mutated" by randomly substituting C residues with G residues, gradually increasing the overall G content of the sequence. This mutational trajectory will eventually arrive at the G homopolymer. We have computed the change in  $\Delta G$  for sequences along this mutational walk for the AU and CG edges of the simplex. Plotted are the mean values of  $\Delta G$  for six mutational walks at 37°C, N = 100. Of these four end member walks, the walk toward poly-G demonstrates the sharpest transition. The majority of evolved RNA lie in the G+C-rich regime of the purine-manifold.

quence composition collapses to two bases (either AU or CG). Biological functionality presumably diminishes as two-base compositions are approached [9,30]: The advantages of a larger alphabet must at some point outweigh advantages of overall thermodynamic stability. Also, there is experimental evidence demonstrating that RNA function does not always require (and can even be inhibited by) maximal thermodynamic stability (e.g., [50–52]). However, these points are moot if in nature this extreme compositional boundary is unapproached by evolving populations. This may in fact be the case, as extreme compositions are impossible to achieve in the absence of profound mutational biases, especially for longer RNA sequences. All the same, the PMF does not state that functional sequences must avoid all conflicting contacts, only that functional molecules tend to be less frustrated than random sequences. With these caveats in mind, we will assume that selection tends to alter base composition to values that intrinsically support minimally frustrated structures.

### Evolutionary Potential

The expectation that a composition vector,  $c$ , will contain useful sequences is given by its calculated mean frustration,  $\langle F_c \rangle$ . The expectation that it will be sampled by mutation is given by its calculated Shannon entropy,  $S_c$ . Selection tends to direct populations toward composition vectors of minimal frustration. Mutation tends to direct populations toward composition vectors of maximal entropy. As defined here, these two forces act in opposition: selection toward the endpoints of Chargaff's Axis, mutation toward the isoheteropolymers. The evolutionary potential of a composition vector  $U_c$  is simply the result of these two forces and can be calculated as:

$$U_c = \langle F_c \rangle - \mu S_c \quad (5)$$

$\mu$  is a constant that adjusts the relative contribution of mutation and selection to evolutionary potential. Evolution will tend to alter the location of the mean base composition of evolving populations to composition vectors of smaller  $U$  values. Superficially, this equation resembles the calculation for free energy in chemical systems, where  $\langle F \rangle$  is an "evolutionary enthalpy" that is minimized via selection, and  $S$  is an entropy term that tends to increase via mutation. In this sense,  $\mu$  is analogous to temperature, and for low values, selection dominates the trajectories of evolving populations as they become centered near the endpoint(s) of Chargaff's Axis. Folded structures become well ordered as they become well adapted. When  $\mu = 0$ ,  $U = \langle F \rangle$  and evolutionary potential is equivalent to the self-organization of the composition vector. For high values of  $\mu$ , mutation will dominate the course of evolutionary trajectories and sequences will become disordered, centered around the iso-

FIGURE 4

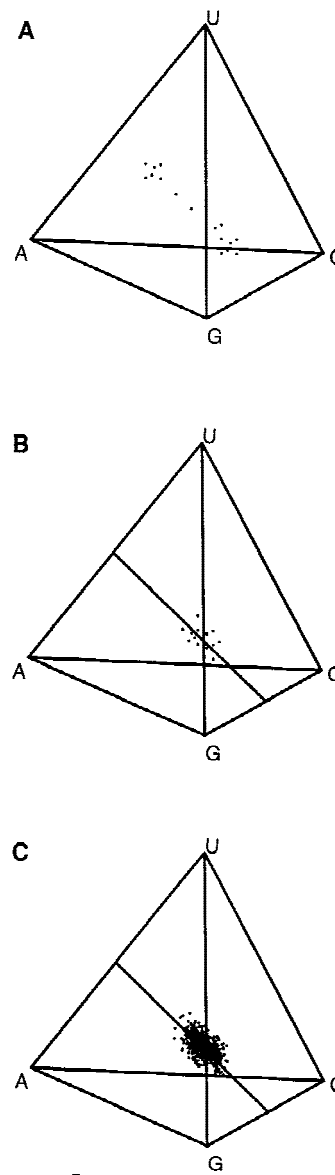


FIGURE 4. Distributions of the top 1 percent evolutionary potential vectors. These vectors represent the compositions most likely to produce functional sequences under random search of sequence space, given the constraints of self-organization. These composition vectors act as dynamical basins of attraction to the base composition of populations of evolving sequences. (A) The distribution of optimal potential vectors for the mean-field calculation has two local optima, each located on Chargaff's Axis on opposite sides of the isoheteropolymers. The distribution is symmetric about Chargaff's Axis. (B) The distribution of optimal potential vectors for the thermodynamic approximation has a global optimum located in a G+C, G+A, and G+U biased province of the simplex. (C) The distribution of 928 cytoplasmic tRNA sequences derived from organisms spanning the universal tree of life. The distribution of 5S rRNA is similar and is therefore not shown. Note the resemblance of this distribution to the optimal potential vectors in B.

**FIGURE 5**

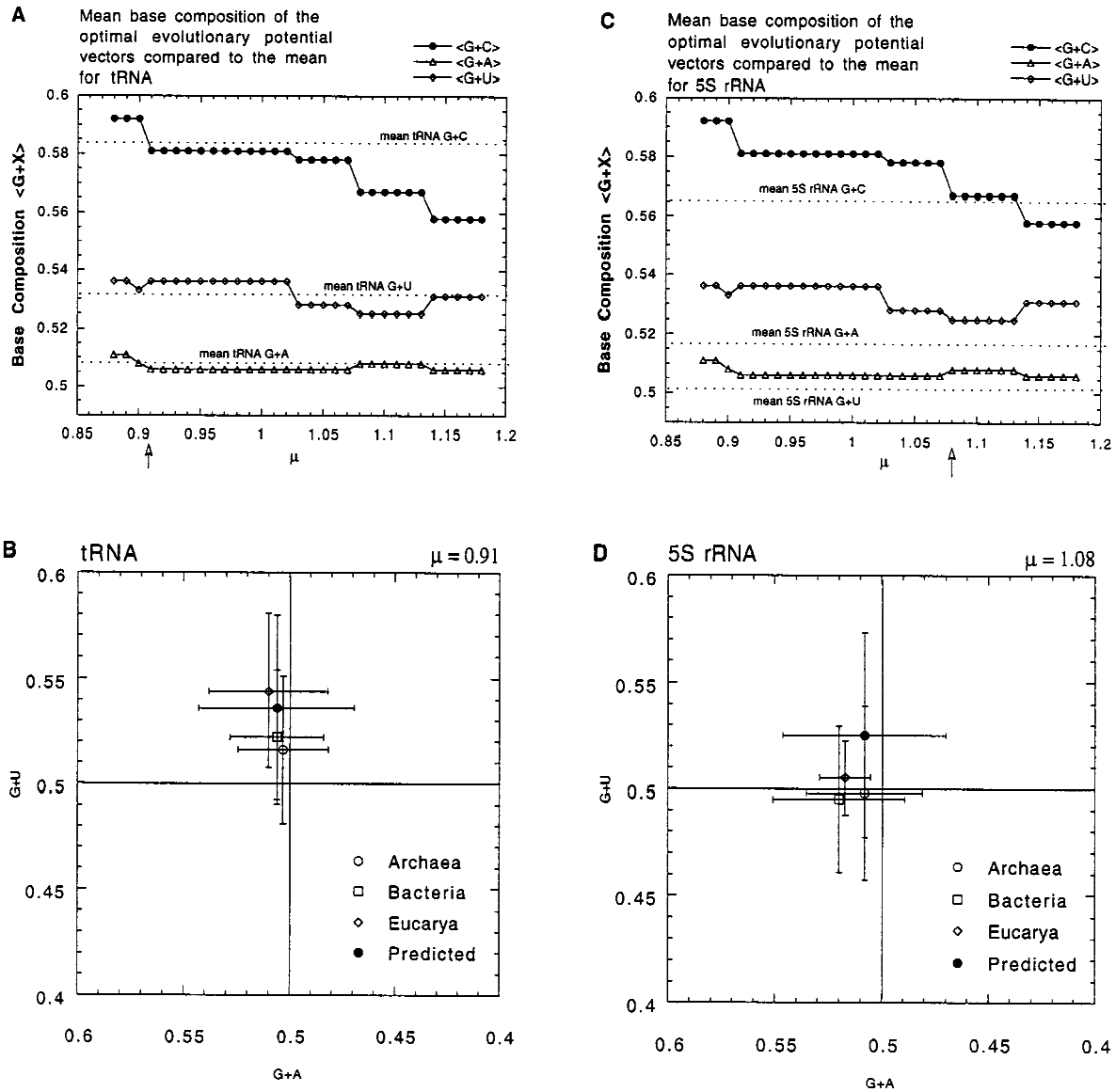


FIGURE 5. Scaling of the evolutionary potential models to actual RNA data. For the range of  $\mu$  values depicted in (A) and (C), the G+C, G+A, and G+U contents remain above 0.5, which is consistent with the compositional biases observed for these and other functional classes of RNA. (A) The mean G+C content of the optimal potential vectors first coincides with the mean G+C content of 928 cytoplasmic tRNA data at  $\mu = 0.91$  (arrow). The predicted mean G+A and G+U values are both consistent with the mean tRNA G+A and G+U contents and are depicted in (B), an analytical projection of the RNA simplex from the CG edge along Chargaff's Axis (center). The distributions of the tRNA are depicted for each phylogenetic domain. Bars represent one standard deviation. (C) The mean G+C content of the optimal potential vectors first coincides with the mean G+C content of 382 5S rRNA data at  $\mu = 1.08$  (arrow). In this case the predicted mean G+A and G+U values are less consistent with the mean G+A and G+U contents of the 5S rRNA data. (D) Projection of the mean G+A and G+U contents of 382 5S rRNA compared to the mean G+A and G+U contents of predicted optimal evolutionary potential vectors.

heteropolymers. It is important to note that evolutionary potential is a statistical property and does not preclude the evolutionary realization of sequences from less optimal composition vectors.

In order to visualize basins of optimal evolutionary potential, we plotted only the top 18 (top 1 percent) optimal potential vectors. We then compared this distribution to the distributions of the cytoplasmic tRNA or 5S rRNA data sets.

Figure 4A depicts the optimal potential vectors for the mean-field approximation with  $\mu$  adjusted to 0.64. At this  $\mu$  value, the optimal potential vectors form a dumbbell-shaped basin of attraction, lying along Chargaff's Axis. The lobes of the dumbbell correspond to the two equally minimal  $\langle F \rangle$  optima derived from the mean-field model. Despite the simplicity of this approximation, the mean-field potential vectors nonetheless indicate the importance of Chargaff's Axis and values of G+C content between the extremes of 0.2 and 0.8 (similar to evolved sequences). For large values of  $\mu$ , the two lobes will converge to form a single symmetrical sphere centered on the isoheteropolymers.

In contrast, the top 1 percent optimal potential vectors derived from the thermodynamic approximation form a single lobe-like basin of attraction (Figure 4B), as the thermodynamic approximation admits only a single optimal  $\langle F \rangle$  value. For a broad range of  $\mu$  values, the overall shape of the optimal potential vectors is remarkably similar to the shape of both the tRNA and 5S rRNA distributions (Figure 4C). Because mean G+C values are most sensitive to changes in  $\mu$ , we adjusted  $\mu$  from low values to higher values until the mean G+C contents of the optimal evolutionary potential vectors first coincide with the mean G+C content of the evolved sequences. We then compared the mean G+A and G+U values of the simulated and empirical data. The fit between the mean G+C content of the optimal potential vectors ( $0.581 \pm 0.0604$ ) and the tRNA ( $0.584 \pm 0.0514$ ) first becomes optimal at  $\mu = 0.91$  (Figure 5A). For this value of  $\mu$ , the optimal potential vectors are slightly G+A and G+U biased with mean values of  $0.506 \pm 0.0369$  (tRNA,  $0.508 \pm 0.0252$ ) and  $0.536 \pm 0.0435$  (tRNA,  $0.531 \pm 0.0365$ ), respectively (Figure 5B). When the same fitting technique is applied to the 5S rRNA, the mean G+C values coincide at  $\mu = 1.08$  (Figure 5C). Though the mean G+A and G+U contents of the optimal potential vectors are similar to the empirical data, the fits are not as close as the in tRNA data set (Figure 5D).

The extraordinary fit to tRNA may reflect the autonomous nature of the molecule under *in vivo* conditions. RNAs such as 5S rRNA and longer molecules are associated with proteins, other RNA, and cofactors under functional conditions. How these intimate intermolecular associations may alter the folding and stability requirements of these molecules and their compositional balances remains an open question. The optimal potential vectors can be simi-

larly compared to other functional classes of RNA. A cursory analysis shows that longer RNA sequences (such as the 3,000 nucleotide 23S rRNA) demonstrate approximately 10 percent higher G+A values than those of tRNA and 5S rRNA. Thus the fits of the simulation to longer RNAs are correspondingly poorer. Refinements or alternative formulations of the selection and mutation terms will be necessary. For example, using computer simulations, Fontana and colleagues [9] have shown that the base composition of RNA sequences provides a means of tuning the structure of adaptive landscapes and might therefore influence evolutionary optimization. The frustration term in Equation 5 could be replaced with a measure of the correlation length on the landscape, and a new vector field for evolutionary potential could then be computed. The ability to compare simulations such as these with empirical sequence data taken from naturally or artificially evolved RNA makes the RNA simplex a useful tool for testing hypotheses concerning self-organization and selection in RNA evolution.

## CONCLUSION

Our framework and simulations propose a mechanism accounting for invariant G+A and G+U contents among disparate RNAs despite drastic evolutionary divergence at the sequence level. We demonstrate that an anisotropy in the underlying distribution of thermodynamic stability in RNA sequence space acts as an attractor, constraining the evolutionary trajectories of RNA to time-invariant, localized provinces of base composition. This underlying distribution of self-organization in RNA illuminates the intrinsic physicochemical constraints that limit the distribution of forms seen in nature.

## ACKNOWLEDGMENTS

We thank J.W. Schopf, C. Marshall, D. Kenan, and Wendy Johnston for helpful discussions and advice. This work was supported by the University of California Institute of Geophysics and Planetary Physics' Center for the Study of the Evolution and Origin of Life, Diversity Biotechnology Consortium, and the Combinatorial Sciences Center at Duke University. Computer simulations were completed at the Santa Fe Institute with support of the Office of Naval Research acting in cooperation with the Defense Advanced Research Project Agency.

## REFERENCES

1. Kauffman, S.A. *The Origins of Order, Self-Organization and Selection in Evolution*. Oxford University Press, Oxford, 1993.
2. Kauffman, S.A. *At Home in the Universe*. Oxford University Press, Oxford, p. 8, 1995.
3. Schultes, E.; Hraber, P.T.; LaBean, T.H. Estimating the Contributions of Selection and Self-Organization in RNA Secondary Structure. *J Mol Mol*, in press.

4. Schultes, E.; Hrabec, P.T.; LaBean, T.H. Global Similarities in Nucleotide Base Composition Among Disparate Functional Classes of Single-Stranded RNA Imply Adaptive Evolutionary Convergence. *RNA* 1997, 3, 792–806.
5. Cairns-Smith, G. *The Life Puzzle*. Oliver and Boyd, Edinburgh, pp. 82–109, 1971.
6. Rechenberg, I. *Evolutionsstrategie, Optimierung Technischer Systeme Nach Prinzipien der Biologischen Evolution*. Stuttgart-Bad Cannstatt, 1973.
7. Smith, J.M. Natural Selection and the Concept of a Protein Space. *Nature* 1970, 225, 563–564.
8. Eigen, M. *Steps Towards Life, A Perspective on Evolution*. Oxford University Press, Oxford, pp. 92–100, 1992.
9. Fontana, W., et al. Statistics of Landscapes Based on Free Energies, Replication and Degradation Rate Constants of RNA Secondary Structures. *Mh Chem* 1991, 122, 795–819.
10. Fontana, W.; Konings, D.A.M.; Stadler, P.F.; Schuster, P. Statistics of RNA Secondary Structures. *Biopolymers* 1993, 33, 1389–1404.
11. Fontana, W.; Schuster, P. Continuity in Evolution: On the Nature of Transitions. *Science* 1998, 280, 1451–1455.
12. Hofacker, I.L., et al. Fast Folding and Comparison of RNA Secondary Structures. *Monatshefte für Chemie* 1994, 125, 167–188.
13. Huynen, M.A.; Hogeweg, P. Pattern Generation in Molecular Evolution: Exploitation of the Variation in RNA Landscapes. *J Mol Evol* 1994, 39, 71–79.
14. Huynen, M.A.; Stadler, P.F.; Fontana, W. Smoothness Within Ruggedness: The Role of Neutrality in Adaptation. *Proc Natl Acad Sci USA* 1996, 93, 397–401.
15. Huynen, M.A.; Gutell, R.; Konings, D. Assessing the Reliability of RNA Folding Using Statistical Mechanics. *J Molec Biol* 1997, 267, 1104–1112.
16. Stadler, P.F.; Gruener, W. Anisotropies in Fitness Landscapes. *J Theor Biol* 1993, 165, 378–388.
17. Tacker, M.; Fontana, W.; Stadler, P.F.; Schuster, P. Statistics of RNA Melting Kinetics. *Eur Biophys J* 1994, 23, 29–38.
18. Langton, C.G. *Computation at the Edge of Chaos: Phase-Transitions and Emergent Computation*. Doctoral dissertation, University of Michigan, pp. 27–31, 1991.
19. Li, W.; Packard, N.H.; Langton, C.G. Transition Phenomena in Cellular Automata Rule Space. *Physica D* 1990, 45, 77–94.
20. Richardson, D.C.; Richardson, J.S. The Kinemage: A Tool for Scientific Communication. *Protein Science* 1992, 1, 3–9.
21. Visualization software and accompanying data files used in this study can be obtained at <http://www.santafe.edu/~pth/simplex.html>.
22. Watson, J.D.; Crick, F.H.C. Molecular Structure of Nucleic Acids. *Nature* 1953, 171, 737–738.
23. Cantor, C.; Schimmel, P. *Biophysical Chemistry, Part III*. W.H. Freeman and Co., New York, p. 162, 1980.
24. Shannon, C.; Weaver, W. *Mathematical Theory of Communication*. University of Illinois Press, Urbana, 1949.
25. Batey, R.T.; Doudna, J.A. The Parallel Universe of RNA Folding. *Nat Struct Biol* 1998, 5, 337–340.
26. Draper, D.E. The RNA-folding Problem. *Acc Chem Res* 1992, 25, 201–207.
27. Draper, D.E. Strategies for RNA Folding. *Trends Biochem Sci* 1996, 21, 165–169.
28. Price, S.; Nagai, K. Secrets of RNA Folding Revealed. *Structure* 1996, 4, 1129–1132.
29. Zarrinkar, P.P.; Williamson, J.R. The Kinetic Folding Pathway of the Tetrahymena Ribozyme Reveal Possible Similarities Between RNA and Protein Folding. *Nat Struct Bio* 1996, 3, 432–438.
30. Brown, J.W.; Hass, E.S.; Pace, N.R. Characterization of Ribonuclease P RNAs From Thermophilic Bacteria. *Nucleic Acids Res* 1993, 21, 671–679.
31. Higgs, P.G. RNA Secondary Structure: A Comparison of Real and Random Sequences. *J Phys I France* 1993, 3, 43–59.
32. Higgs, P.G. Thermodynamic Properties of Transfer-RNA—A Computational Study. *J Chem Soc Faraday T* 1995, 91, 2531–2540.
33. Puglisi J.D.; Tinoco, I. Absorbance Melting Curves of RNA, in *Methods in Enzymology*, Vol. 180, Academic Press, pp. 204–325, 1989.
34. Bryngelson, J.D.; Onuchic, J.N.; Succi, N.D.; Wolynes, P.G. Funnels, Pathways and the Energy Landscape of Protein Folding: A Synthesis. *Proteins* 1995, 21, 167–195.
35. Frauenfelder, H.; Wolynes, P.G.: Biomolecules: Where the Physics of Complexity and Simplicity Meet. *Physics Today* 1994, 47, 58–64.
36. Frauenfelder, H. Complexity: Metaphores, Models and Reality, Santa Fe Institute Studies in the Sciences of Complexity, XIX. G. Cowan, D. Pines, and D. Meltzer (Eds.) Addison-Wesley, Reading, MA, pp. 179–180, 1994.
37. Wolynes, P.G.; Onuchic, J.N.; Thirumalai, D. Navigating the Folding Routes. *Science* 1995, 267, 1619–1620.
38. McCaskill, J.S. The Equilibrium Partition Function and Base Pair Binding Probabilities for RNA Secondary Structure. *Biopolymers* 1990, 29, 1105–1119.
39. Turner D.H.; Sugimoto, N. RNA Structure Prediction. *Ann Rev Biophys Chem* 1988, 17, 167–192.
40. Zuker, M.; Stiegler, P. Optimal Computer Folding of Large RNA Sequences Using Thermodynamics and Auxiliary Information. *Nucleic Acids Res* 1981, 9, 133–149.
41. Zuker, M. On Finding All Suboptimal Folding of an RNA Molecule. *Science* 1989, 244, 48–52.
42. Saenger, W. *Principles of Nucleic Acid Structure*. Springer-Verlag, New York, p. 135, 1984.
43. Blake, R.D.; Hess, S.T.; Nicholson-Tuell, J. The Influence of Nearest Neighbors on the Rate and Pattern of Spontaneous Point Mutations. *J Mol Evol* 1992, 34, 189–200.
44. Gu, X.; Li, W. A Model for the Correlation of Mutation Rate With GC Content and the Origin of GC-Rich Isochores. *J Mol Evol* 1994, 38, 468–475.
45. Li, W.; Wu, C.; Luo, C. Nonrandomness of Point Mutation as Reflected in the Nucleotide Substitutions in Pseudogenes and Its Evolutionary Implications. *J Mol Evol* 1984, 21, 58–71.
46. Pearl, L.H.; Savva, R. The Problem With Pyrimidines. *Nat Struct Biol* 1996, 3, 485–487.
47. Fleischmann, R.D., et al. Whole-Genome Random Sequencing and Assembly of *Haemophilus influenzae*. *Rd Science* 1995, 269, 496–512.
48. Fraser, C.M., et al. The Minimal Gene Complement of *Mycoplasma genitalium*. *Science* 1995, 270, 397–403.
49. Rejto, P.A.; Verkhivker, G.M. Unraveling Principles of Lead Discovery: From Unfrustrated Energy Landscapes to Novel Molecular Anchors. *Proc Natl Acad Sci USA* 1996, 93, 8945–8950.
50. Honda, M.; Brown, E.A.; Lemon, S.M. Stability of a Stem-Loop Involving the Initiator AUG Controls the Efficiency of Initiation of Translation on Hepatitis C Virus RNA. *RNA* 1996, 2, 955–968.
51. Klovins, J.; van Duin, J.; Olsthoorn, R.C. Rescue of the RNA Phage Genome From RNase III Cleavage. *Nucleic Acids Res* 1997, 25, 4201–4208.
52. Olsthoorn, R.C.; Licis, N.; van Duin, J. Leeway and Constraints in the Forced Evolution of a Regulatory RNA Helix. *EMBO J* 1994, 13, 2660–2668.