

## Estimating the Contributions of Selection and Self-Organization in RNA Secondary Structure

Erik A. Schultes,<sup>1,2,\*</sup> Peter T. Hraber,<sup>3,4</sup> Thomas H. LaBean<sup>2,5</sup>

<sup>1</sup> Department of Earth and Space Sciences, University of California at Los Angeles, Los Angeles, CA 90024, USA

<sup>2</sup> Combinatorial Sciences Center, Duke University Medical Center, Durham, NC 27710, USA

<sup>3</sup> Department of Biology, University of New Mexico, Albuquerque, NM 87131, USA

<sup>4</sup> National Center for Genome Resources, 1800 Old Pecos Trail, Suite A, Santa Fe, NM 87501, USA

<sup>5</sup> Department of Biochemistry, Duke University Medical Center, Durham, NC 27710, USA

Received: 25 November 1998 / Accepted: 12 February 1999

**Abstract.** In addition to characteristic structural properties imposed by evolutionary modification, evolved, single-stranded RNAs also display characteristic structural properties imposed by intrinsic physical constraints on RNA polymer folding. The balance of intrinsic and functionally selected characters in the folded conformation of evolved secondary structures was determined by comparing the predicted secondary structures of evolved and unevolved (random) RNA sequences. Though evolved conformations are significantly more ordered than conformations of random-sequence RNA, this analysis demonstrates that the majority of conformational order within evolved structures results not from evolutionary optimization but from constraints imposed by rules intrinsic to RNA polymer folding.

**Key words:** RNA structure — Morphospace — Phenetics — Intrinsic evolutionary constraints — Extrinsic evolutionary constraints

### Introduction

The contribution of intrinsic and extrinsic constraints to the evolution of biological form remains one of the most

important, yet least explored questions in evolutionary biology (Thompson 1917; Seilacher 1991; Raff 1996). The most effective means of resolving the relative importance of these constraints is to compare the distribution of evolved morphologies to the larger space of unevolved possibilities. Such a phenetic (in contrast to cladistic) analysis is based on an appropriate “morphospace” in which disparately related forms can be conveniently compared. However, the formulation of developmentally meaningful and taxonomically comprehensive morphospaces has remained elusive (Gould 1991). Single-stranded RNA provides the ideal system in which to experimentally investigate the role of intrinsic evolutionary constraints as (i) both evolved and unevolved (i.e., random) RNA sequences can be synthesized and their folded conformations probed and compared, (ii) large sequence databases of disparately related molecules currently exist (Schultes et al. 1997), and (iii) RNA can be evolved *in vitro*, enabling repeated evolutionary experiments (Lato et al. 1995). Though numerous theoretical (e.g., Huynen et al. 1996; Fontana and Schuster 1998) and experimental (e.g., Jeager 1997; Unaru and Bartel 1998) studies have focused on the origin and diversification of *functional* RNA, the evolutionary constraints imposed by structural properties intrinsic to RNA polymer chemistry remain poorly characterized. Single-stranded RNA molecules with structure dependent functions are known to possess well-ordered conformations that are both thermodynamically stable and uniquely folded (Draper 1996). Are these properties common or

\*Present address: Whitehead Institute for Biomedical Research, 9 Cambridge Center, Cambridge, MA 02142, USA

Correspondence to: Erik Schultes; e-mail: schultes@wi.mit.edu

rare among random heteropolymers? How hard must selection search sequence space to access RNA having suitably ordered conformations? In what sense have the nucleotide sequences of evolved RNA been informed by natural selection?

To address these evolutionary questions, we have applied recent results from the “new view” of protein folding kinetics to the RNA folding problem (Dill and Chan 1997). Macromolecular folding is driven by the formation of low-energy, intramolecular contacts. In sufficiently complicated polymers the formation of one favorable contact may disrupt, or “frustrate,” the formation of other favorable contacts, delaying or preventing the acquisition of a stable and unique fold. Even the native configuration of a well-adapted sequence may contain residual high-energy contacts which are unavoidable given the complexity of the interactions throughout the molecule. Overall frustration can be reduced only by judiciously altering the sequence of nucleotide bases such that high-energy contacts are avoided, while the native configuration remains intact. Though this represents an extremely complicated combinatorial optimization problem from the standpoint of rational design, we predict that functional RNA sequences have evolved to minimize frustrated intramolecular interactions with respect to random heteropolymer sequences. This “principle of minimal frustration” (PMF) (Bryngelson et al. 1995) is a generic response of RNA to any selection pressure requiring well-ordered structure. Here, we used standard energy minimization algorithms to compare properties of calculated minimum free energy secondary structures of RNA sequences derived from biological specimens to those of analogous random sequences (same length and base composition).

## Methods

*Defining a Molecular Morphospace.* To define quantitatively the stability and uniqueness of RNA secondary structure, we hypothesize that the evolution of functional RNA sequences from random-sequence populations (at constant conditions) is analogous to the adaptation of RNA sequences from mesothermal to hyperthermal conditions, as noted by Brown et al. (1993). Specifically, the evolution of RNA molecules (from random sequences) requiring precisely folded conformations will (i) increase the number of hydrogen bonds in stem helical regions, (ii) reduce irregularities in stem helices, (iii) increase the number of Watson–Crick base pairs at the base of stem loops, (iv) shorten connections between helices, and (v) reduce the number of alternative, stable, conformations. Evolutionary modifications i–iv are effectively summarized by measuring two features of the secondary structure: the base-pairing propensity,  $P$  (measured as the number of base pairs normalized to the length of the sequence in nucleotide residues,  $N$ ), and the mean length of helical stem structures,  $S$  (measured as base pairs), within individual sequences.

Values for  $P$  range from no base-pair interactions,  $P_{\min} = 0.0$ , to an ideal maximum number of base-pair interactions,  $P_{\max} = 0.5$  (in this case, all  $N$  bases pair, yielding  $N/2$  base pairs). Though the RNA folding algorithm can express the thermodynamic stability of folded conformations as the free energy of folding (kcal/mol), base-pairing

propensity was used because  $P$  (i) normalizes for sequence length, (ii) controls for differences in bond strength between A · U, C · G, and G · U base pairs, and (iii) enables direct comparison between simulated secondary structures and secondary structures inferred from sequence comparison studies (Fontana et al. 1993). A contiguous helical stem structure is defined as the number of uninterrupted Watson–Crick and/or G · U bases pairs. Hence, bulges, loops, and unpaired bases terminate helical stem structures. Values for  $S$  range from  $S_{\min} = 1.0$  for single base-paired structures to an idealized hairpin with no loop structure,  $S_{\max} = N/2$ .

Reducing the number of different folded conformations having comparable stability (evolutionary modification  $\nu$  from above) concerns not only the minimum free energy structure but its relation to competing, suboptimal folds. Functional sequences must fold to unique, as well as stable conformations (Herschlag 1995). We have developed an index of the uniqueness of the folded conformation,  $Q$ , determined from the base-pairing probability distribution of the calculated minimum free energy secondary structure. In addition to calculating a minimum-free energy secondary structure, the RNA folding algorithm calculates base-pairing probabilities for each possible nucleotide pair in the sequence (the so-called equilibrium partition function or base-pairing probability matrix). Since these probabilities are based on thermodynamic likelihoods of base-pair formation, the partition function is a summary of the many possible alternative conformations that compete during the energy minimization process of polymer folding. The base-pairing probability matrix has two extreme possibilities: each base-pair combination may be equally likely (with no single secondary structure preferred) or a single base-pairing combination may be specified with no uncertainty (either paired or unpaired) specifying a single, well-defined secondary structure. In practice, intermediate cases prevail. We calculate the uncertainty in base-pairing and therefore secondary structure as the Shannon entropy of the base-pairing probability matrix normalized to sequence length [a slightly different measure was independently developed by Huynen et al. (1997)].

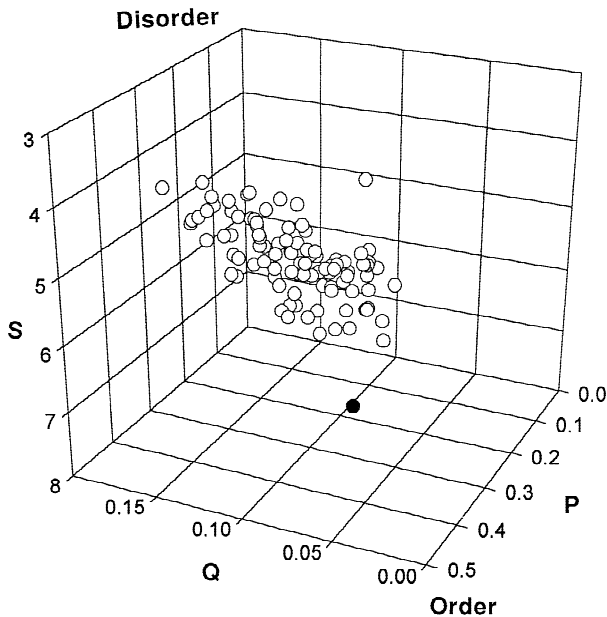
$$Q = -\frac{1}{Q_{\max}} \sum_{i=1}^{N-1} \sum_{j>i}^N p_{ij} \log_2 p_{ij} \quad (1)$$

where

$$Q_{\max} = \frac{1}{2} N \log_2 N \quad (2)$$

A sequence having a well-defined secondary structure has a partition function with low informational entropy. A sequence having numerous conflicting interactions (high frustration) has more uncertainty and higher informational entropy in the partition function. Values for  $Q$  range from  $Q = 0.0$  for perfectly defined, unique structures to  $Q = 1.0$  for sequences having no preferred structure. Taken together,  $P$ ,  $S$ , and  $Q$  define the degree of conformational order of the minimum-free energy secondary structure as a point in a three-dimensional, order-disorder molecular “morphospace” (Fig. 1).

*RNA Folding Algorithm.* We have employed RNAfold, a dynamic programming search method for finding minimal-free energy secondary structures that is based on empirically derived thermodynamic parameters (Zuker and Steigler 1981). We obtained software as the Vienna RNA Package (Hofacker et al. 1994): <ftp://ftp.iti.univie.ac.at/pub/RNA/ViennaRNA-1.1b>. Minimum-free energy structures were calculated at 37°C. It is important to note that there exist significant



**Fig. 1.**  $P$ - $S$ - $Q$  define a comprehensive morphospace for RNA secondary structure where the disparity among the folded conformations of related, unrelated, and even random sequences can be meaningfully quantified. Well-ordered structures have large  $P$  and  $S$  values (a large proportion of base pairs and long stem lengths) and small  $Q$  values (unique structures). The calculated  $P$ ,  $S$ , and  $Q$  values of the *Sulfolobus acidocaldarius* P RNA (filled circle) and those of the permuted cohort (open circles) occupy separate localities of morphospace confirming a preferred direction to evolutionary modification regardless of the disparate functions and evolutionary histories of the molecules investigated. Selection tends to modify the RNA structure in the direction from the disorder regime toward the order regime (bottom of figure). The permuted cohort distributions examined here have similar shapes and locations in morphospace regardless of sequence length.

discrepancies between the computational predictions of minimum-free energy secondary structures and those of secondary structures inferred from sequence comparison studies (Fontana et al. 1993; Huynen et al. 1997; Fontana and Schuster 1998). In general, we find that the structures predicted from simulations have higher  $P$  and  $S$  values than secondary structures inferred from sequence comparison studies (see bracketed values in Table 1). Though it is currently impossible to predict reliably the exact conformation for any particular sequence using energy minimization algorithms, this does not preclude testing for statistical differences between evolved and random sequences. Since we are testing the PMF by comparing two different "samples" of RNA sequences (i.e., comparing evolved sequences to unevolved sequences) using the same algorithm, inaccuracies in the predicted structures are tolerated as systematic error. We stress that the focus of this investigation is not the prediction of exact secondary structure from sequence information but is instead the statistical analysis of the effects of sequence evolution on RNA structure. The model independence gained from statistical comparison of evolved and unevolved RNA sequences contrasts to previous numerical studies concerning the PMF in protein sequences where the simplified approximations of protein folding do not admit the use of authentic, naturally evolved protein sequence data (Shakhnovich and Gutin 1990; Li et al. 1996).

**Statistical Analysis and Phylogenetic Controls.** Using Student's  $t$  test as an inferential statistic, we examined the following hypotheses to test whether the PMF holds for natural sequences: evolved sequences have a greater number of base pairs (higher  $P$  value), longer mean stem lengths (higher  $S$  values), and more unique minimum-free energy struc-

tures (lower  $Q$  values) than random, unevolved sequences. We then compiled a database of 24, full-length, naturally evolved RNA sequences (Table 1). These disparate sequences represent six distinct functional classes and the three phylogenetic domains of the universal tree of life. mRNAs were not included in this analysis, as their structural requirements are less well understood and, as such, lie beyond the scope of this study. Sequences belonging to disparate functional classes share no sequence similarity (above random expectations) and therefore no interpretable evolutionary history. Hence, similarities among  $P$ ,  $S$ , and  $Q$  values between functional classes must be due to adaptive convergent evolution rather than shared ancestry (Schultes et al. 1997). Further, as an independent phylogenetic control, we also include seven artificially evolved self-ligating ribozymes isolated *in vitro* from synthetic random sequence libraries (Bartel and Szostak 1993). Because artificial ribozymes are known to be genealogically independent of biological lineages, any similarities in conformational properties are interpreted as adaptive convergent evolution and not historical constraints. The seven artificial, self-ligating RNAs belong to three disparate classes that were evolved independently from three disparate random sequences sharing no sequence similarity (Eklund et al. 1995). Our database therefore contains 31 sequences representing up to nine independent evolutionary lineages.

**Random-Sequence Generation.** The calculated secondary structures of the 31 evolved RNA sequences were compared to those of randomly generated sequences. Controlling for base composition among random sequences is essential, as base composition is known to influence the thermodynamic stability of nucleic acids (Saenger 1984). Hence, for each of the 31 RNAs, a population of 100 random heteropolymers was constructed by permuting the evolved sequence. The permutation procedure consisted of three perfect shuffles, where a perfect shuffle sequentially swaps nucleotides at all sites with a randomly chosen site elsewhere in the sequence (Knuth 1973). We refer to this population as the permuted cohort of the original RNA. Though permutations effectively randomize the primary structure (and therefore secondary structure), they leave the base composition and sequence length unchanged. The mean values of  $P$ ,  $S$ , and  $Q$  for each permuted cohort population were then compared to the corresponding natural sequence to test the hypotheses listed above.

**Database Construction.** 5S rRNA sequences were obtained from <http://cammsg3.caos.kun.nl>. tRNA sequences were obtained from <ftp://ftp.ebi.ac.uk/pub/databases/trna>. P RNA sequences were obtained from the Ribonuclease P Database. Artificial self-ligating ribozymes and their GenBank files are described by Bartel and Szostak (1993) and Eklund et al. (1995). Organisms in Table 1 are from, top to bottom, *H. morrhuae*, *H. trapanicum*, *S. acidocaldarius*, *P. woesei*, *P. yezoensis*, *M. voltae*, *U. scabiosae*, *H. cutirubrum*, *T. tenax*, *S. PCC6301*, *Syn-echocystis* sp. (U10482), *R. rubrum*, *E. coli*, *M. fermentans*, *A. nidulans*, *P. cepacia*, *E. coli*, *C. ellipsoidea* (X63520), *T. thermophila* (J01235), *X. laevis*, *S. obliquus* (X17375), bacteriophage SP82 (U04812), chloroella virus (D29631), and phage T5.

## Results

### *The Relationship Between Evolved and Unevolved RNA Secondary Structures*

The results detailed in Table 1 demonstrate that evolved and unevolved sequences, unrelated by sequence similarity, nonetheless have remarkably similar values for  $P$ ,  $S$ , and  $Q$ . For example, the calculated  $P$  values among the 31 evolved RNA sequences in Table 1 average to 0.305

**Table 1.** Conformational order measured as  $P$ ,  $Q$ , and  $S$  among evolved ssRNA in relation to their corresponding permuted cohort<sup>a</sup>

RNA	$N$	$P_e$	$\langle P_{\text{coh}} \rangle$	$Q_e$	$\langle Q_{\text{coh}} \rangle$	$S_e$	$\langle S_{\text{coh}} \rangle$
Archaea							
P RNA	476	0.319	$0.306 \pm 0.0090$ (94) <sup>b</sup>	0.060	$0.090 \pm 0.0281$ (86) <sup>b</sup>	4.83	$4.60 \pm 0.271$ (77) <sup>b</sup>
P RNA	358	0.310	$0.288 \pm 0.0122$ (98) <sup>b</sup>	0.067	$0.098 \pm 0.0290$ (83) <sup>b</sup>	4.63	$4.37 \pm 0.287$ (79) <sup>b</sup>
P RNA	315	0.349 [0.302]	$0.286 \pm 0.0186$ (100) <sup>b</sup>	0.066	$0.110 \pm 0.316$ (93) <sup>b</sup>	6.88 [7.18]	$5.17 \pm 0.454$ (100) <sup>b</sup>
5S rRNA	124	0.331	$0.301 \pm 0.0252$ (91) <sup>b</sup>	0.029	$0.112 \pm 0.0405$ (99) <sup>b</sup>	4.56	$4.90 \pm 0.613$ (33)
5S rRNA	121	0.298	$0.256 \pm 0.0315$ (92) <sup>b</sup>	0.085	$0.103 \pm 0.0318$ (68) <sup>b</sup>	5.54	$4.53 \pm 0.586$ (94) <sup>b</sup>
5S rRNA	120	0.242	$0.253 \pm 0.0282$ (37) <sup>b</sup>	0.084	$0.108 \pm 0.0379$ (70) <sup>b</sup>	4.46	$4.83 \pm 0.605$ (31)
5S rRNA	119	0.336	$0.285 \pm 0.0296$ (97) <sup>b</sup>	0.047	$0.102 \pm 0.0383$ (96) <sup>b</sup>	5.33	$4.83 \pm 0.738$ (81) <sup>b</sup>
tRNA-GCA	76	0.276	$0.284 \pm 0.0312$ (45)	0.119	$0.106 \pm 0.0472$ (38)	6.00	$4.91 \pm 0.755$ (93) <sup>b</sup>
tRNA-CGC	72	0.292	$0.301 \pm 0.0269$ (45)	0.095	$0.106 \pm 0.0483$ (54)	6.00	$4.91 \pm 0.817$ (93) <sup>b</sup>
Bacteria							
16S rRNA	1487	0.330 [0.301]	$0.320 \pm 0.0060$ (93) <sup>b</sup>	0.076	$0.092 \pm 0.0194$ (80) <sup>b</sup>	5.54 [4.14]	$5.28 \pm 0.177$ (92) <sup>b</sup>
Group I	655	0.321	$0.309 \pm 0.0135$ (79) <sup>b</sup>	0.095	$0.107 \pm 0.0241$ (66) <sup>b</sup>	5.32	$5.48 \pm 0.333$ (34)
P RNA	429	0.322	$0.325 \pm 0.0120$ (38)	0.045	$0.102 \pm 0.0284$ (99) <sup>b</sup>	5.31	$5.23 \pm 0.320$ (64) <sup>b</sup>
P RNA	377	0.337 [0.284]	$0.305 \pm 0.0131$ (100) <sup>b</sup>	0.113	$0.099 \pm 0.0286$ (31)	4.98 [4.61]	$4.86 \pm 0.344$ (66) <sup>b</sup>
P RNA	302	0.265	$0.260 \pm 0.0202$ (60)	0.090	$0.109 \pm 0.0268$ (74) <sup>b</sup>	4.44	$4.81 \pm 0.472$ (21)
5S rRNA	120	0.317	$0.264 \pm 0.0308$ (98) <sup>b</sup>	0.084	$0.106 \pm 0.0389$ (69) <sup>b</sup>	4.47	$4.90 \pm 0.605$ (23)
5S rRNA	116	0.267	$0.268 \pm 0.0309$ (56)	0.066	$0.104 \pm 0.0377$ (86) <sup>b</sup>	5.64	$4.88 \pm 0.712$ (84) <sup>b</sup>
tRNA-GGC	76	0.303	$0.279 \pm 0.0301$ (86) <sup>b</sup>	0.098	$0.102 \pm 0.0454$ (51)	6.57	$5.03 \pm 0.934$ (94) <sup>b</sup>
Eucarya							
Group I	441	0.293	$0.303 \pm 0.0144$ (28)	0.073	$0.099 \pm 0.0302$ (80) <sup>b</sup>	5.86	$5.20 \pm 0.346$ (96) <sup>b</sup>
Group I	413	0.334	$0.297 \pm 0.0161$ (99) <sup>b</sup>	0.050	$0.111 \pm 0.0254$ (100) <sup>b</sup>	6.42	$5.35 \pm 0.403$ (98) <sup>b</sup>
tRNA-TGC	69	0.362	$0.237 \pm 0.0545$ (100) <sup>b</sup>	0.118	$0.123 \pm 0.0412$ (59)	5.56	$5.71 \pm 1.294$ (49)
Mitochondrial							
Group II	608	0.308	$0.302 \pm 0.0134$ (66) <sup>b</sup>	0.072	$0.103 \pm 0.0210$ (92) <sup>b</sup>	5.67	$5.32 \pm 0.355$ (100) <sup>b</sup>
Virus							
Group I	915	0.286	$0.270 \pm 0.0132$ (89) <sup>b</sup>	0.078	$0.104 \pm 0.0172$ (93) <sup>b</sup>	5.29	$4.97 \pm 0.258$ (91) <sup>b</sup>
Group I	399	0.303	$0.283 \pm 0.0186$ (83) <sup>b</sup>	0.101	$0.108 \pm 0.0254$ (58)	5.38	$5.20 \pm 0.398$ (68) <sup>b</sup>
tRNA-TGC	75	0.293	$0.275 \pm 0.0349$ (74) <sup>b</sup>	0.046	$0.106 \pm 0.0421$ (93) <sup>b</sup>	6.29	$4.96 \pm 0.858$ (93) <sup>b</sup>
Artificial self-ligating ribozymes							
Class I							
Isolate b1	274	0.296	$0.279 \pm 0.0203$ (81) <sup>b</sup>	0.148	$0.107 \pm 0.0335$ (13)	4.91	$5.22 \pm 0.457$ (26)
Construct b1207	119	0.227	$0.249 \pm 0.0252$ (19)	0.082	$0.101 \pm 0.0356$ (68)	3.86	$4.51 \pm 0.633$ (12)
Class II							
Isolate c2	271	0.328	$0.316 \pm 0.0169$ (80) <sup>b</sup>	0.111	$0.108 \pm 0.0317$ (44)	5.09	$5.37 \pm 0.470$ (27)
Isolate d1	273	0.304	$0.298 \pm 0.0201$ (64) <sup>b</sup>	0.109	$0.113 \pm 0.0349$ (53)	5.93	$5.35 \pm 0.498$ (88) <sup>b</sup>
Isolate f1	273	0.293	$0.299 \pm 0.0192$ (40)	0.129	$0.107 \pm 0.0299$ (19)	4.71	$5.19 \pm 0.524$ (17)
Class III							
Isolate e3	272	0.313	$0.314 \pm 0.0184$ (50)	0.123	$0.111 \pm 0.0310$ (34)	7.73	$5.75 \pm 0.544$ (100) <sup>b</sup>
Isolage g1	274	0.307	$0.303 \pm 0.0203$ (60)	0.082	$0.111 \pm 0.0322$ (81) <sup>b</sup>	5.25	$5.29 \pm 0.472$ (51)

<sup>a</sup>  $N$  is the sequence length in nucleotide residues. Angle braces indicate mean values for the 100 sequences of permuted cohort. Variability is specified as  $\pm 1$  SD. Brackets indicate values derived from inferred secondary structures taken from the literature and the Ribonuclease P Database (<http://www.mbio.ncsu.edu/RNaseP>). Parentheses indicate the percentile rank of the evolved RNA sequence among its permuted cohort.

<sup>b</sup> Statistically significant difference between values of evolved (subscript e) RNA sequences and their corresponding permuted cohort (subscript coh) means at  $p = 0.05$ .

$\pm 0.0289$ , a result similar to a restricted range of  $P$  values ( $0.3026 \pm 0.0269$ ) observed among inferred secondary structures derived from sequence comparison studies [ $P$  values for 10 unique sequences representing five distinct functional classes (23S rRNA, 16S rRNA, 5S rRNA, P RNA, and tRNA) were calculated by Schultes et al. (1997)]. Given the structural and sequence disparity among these different functional classes, the degree of conformational order is nonetheless remarkably consistent.

Even among permuted cohort sequences we find that thermodynamic constraints ensure a characteristic degree of ordered structure. However, sequences that have

evolved (both naturally and artificially) in most cases have a statistically significantly greater thermodynamic stability and uniqueness of secondary structure. Compared to the means of their permuted cohort, 65% of the 31 sequences demonstrate significantly larger  $P$  values, 68% statistically significantly larger  $S$  values, and 61% significantly smaller  $Q$  values. These data corroborate results that were independently obtained in similar studies by Higgs (1993, 1995). Table 1 lists (in parentheses) the percentile rank of each evolved sequence among the permuted cohort distribution. Though it is not surprising that evolved structures have increased stability and uniqueness, these data contrast with results from protein

sequence analyses (using statistical run tests), indicating that amino acid sequences of evolved proteins are indistinguishable from random heteropolymers (White 1994). Apparently, phylogenetically and functionally independent trends toward greater structural stability and uniqueness can nevertheless arise within the unique evolutionary histories of varied biochemical adaptations.

Additionally, in Table 1, there appears to be an important difference between the proportion of significant values between naturally evolved and artificially evolved RNA. Only 28.6% of the  $P$ ,  $S$ , and  $Q$  values of the 7 artificially evolved ligase ribozymes are significantly more ordered than random sequences, in contrast to 83.3% of the  $P$ ,  $S$ , and  $Q$  values of the 24 naturally evolved RNA. This discrepancy may be due to the fact that artificial sequences have experienced a relatively limited evolutionary history (only 10 rounds of *in vitro* selection–amplification) under relatively simple conditions. Presumably, additional rounds of selection–mutation–amplification and/or greater stringency of the selection protocol would increase the proportion of artificial RNA having significantly increased structural order. If subsequences, known to be neutral, are removed from the full-length artificial ribozymes and the resulting, shorter sequences refolded, we find the proportion of significantly different  $P$ ,  $S$ , and  $Q$  values increases from 28.6 to 51.5% [the neutral subsequences are annotated in GenBank files (Bartel and Szostak 1993; Ekland et al. 1995)]. Apparently, neutral subsequences which cannot be removed by existing *in vitro* selection technology tend to disorder the ligase conformation.

#### *Self-Organization: The Spontaneous Minimization of Frustration*

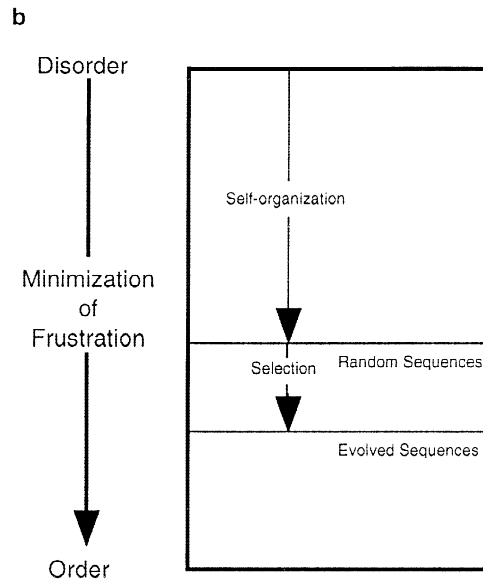
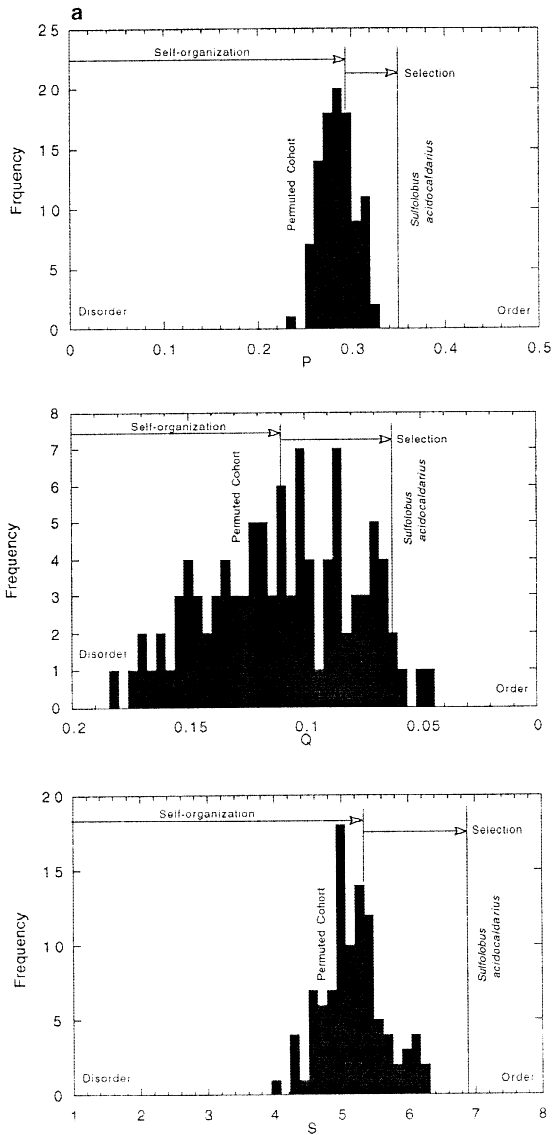
The mean values of  $P$ ,  $S$ , and  $Q$  derived from the permuted cohort populations provide an index of the intrinsic order, or degree of self-organization, among average, random, RNA polymers. Defined as a mean expectation, self-organization is a statistical property of the permuted cohort. Individual sequences may have more or less conformational order than the mean values of the permuted cohort but will be found less frequently via random search. In addition to an expected degree of intrinsic organization, evolved sequences have an extrinsic source of order—that of natural selection. Whereas the PMF states that natural selection minimizes conflicting intramolecular interactions, self-organization is a measure of the degree to which random (i.e., unselected and unevolved) sequences minimize intramolecular frustration spontaneously. Hence, the total conformational order among well-adapted sequences has an intrinsic component (self-organization) and extrinsic component (selection). The difference between the mean values of the permuted cohort population and the values derived from the corresponding evolved sequence represents the

amount of order due to selection (Fig. 2). These differences demonstrate that at the level of secondary structure, more than 90.0% of the conformational order found in the base-pairing and stem lengths for evolved sequences can be attributed to the intrinsic ordering capacity of RNA. Random sequences have a base-pairing propensity ( $P$ ) that averages 0.288, while evolved sequences have an average base-pairing propensity that is slightly greater at 0.318, a difference of only 9.43%. Likewise, stem lengths of evolved RNA structures, averaging 5.70 nucleotides long, are only 12.09% longer on average than those of random sequences. The uniqueness of the folded conformation appears to be most sensitive to the effects of selection: the  $Q$  values of random sequences average 0.104, while evolved structures have an average  $Q$  value of 0.068, an increase in the uniqueness of the conformation of 34.6%. The magnitude of conformational order of random RNA is substantial and appears, in roughly 30–40% of the sequences, to be indistinguishable from that of evolved structures. The ordering effects of selection on RNA structure, though in many cases significant, appear to be small in magnitude.

#### **Discussion**

The results presented above indicate that evolutionary processes systematically deform the folding energy landscape of functional RNAs compared to random sequences. However, the majority of the conformational order found in functional RNA appears not to be the result of a long history of evolutionary modification but is inherent in the physiochemical interactions that drive RNA folding. Though selection undoubtedly modifies RNA sequence and structure in the adaptation to specific functions and environments, the majority of conformational order is intrinsic [i.e., comes “for free” (Kauffman 1993)]. The high degree of ordered structure among random RNA sequences is consistent with experimental results which show that large and complex ribozymes can emerge from an extremely limited sampling of sequence possibilities, which in turn implies the existence of a large number of distinct and complex structures throughout sequence space (Ekland and Bartel 1995).

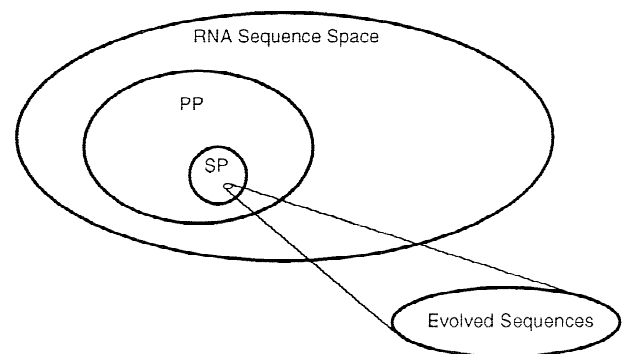
The relationship between random and evolved sequences can therefore be summarized as nested sets of increasing biological relevance (Fig. 3). An individual RNA, such as the phage T5 TGC-tRNA (Table 1), represents an evolved instance of a class of sequences that are functionally equivalent for translational incorporation of cysteine in the context of T5 replication. Sequences belonging to this class have specific adaptations, or a *specific phenotype*, that is idiosyncratic to that particular function. The specific phenotype includes adaptations such as the guide sequence of self-splicing introns, the anticodon stem-loop structure of tRNAs, or



**Fig. 2.** Estimating the relative contribution of self-organization and selection in evolved RNA secondary structures. **a** The  $P$ ,  $S$ , and  $Q$  values of *Sulfolobus acidocaldarius* P RNA and the mean  $P$ ,  $S$ , and  $Q$  values of its corresponding permuted cohort. The mean of the permuted cohort values is a measure of the intrinsic order found among random heteropolymers that have not been modified by evolutionary processes. The difference between the mean value of the permuted cohort and that of the evolved sequence provides a measure of the contribution of self-organization (i.e., the spontaneous minimization of frustration) and selection to the total conformational order of the evolved RNA. **b** A schematic representation of the intrinsic (self-organization) and extrinsic (selection) sources of conformational order found among evolved RNA sequences. The majority of the conformational order necessary to perform biological functions is due to intrinsic organization, and not selection.

possibly the U-rich composition of certain snRNAs. The specific phenotype is unique to individual evolutionary lineages and forms the basis of comparative sequence analysis. All sequences having the specific phenotype belong to a more general class of sequences having the proper *prerequisite phenotype*. These characters are not unique to particular structural adaptations and evolutionary lineages but are, instead, general properties like increased thermodynamic stability and uniqueness of folded conformation that are necessary (but not sufficient) for a wide variety of functions. As such, the prerequisite phenotype can be elucidated only by comparing independently evolved sequences with completely random sequences where the confounding influences of shared ancestry can be eliminated. Our results indicate that, at least for  $P$ ,  $S$ , and  $Q$ , random sequences having biological base compositions often demonstrate the prerequisite phenotype.

Organizing RNA sequence space as in Fig. 3 has theoretical and practical implications. Theoretically, the



**Fig. 3.** The relationship between random and evolved sequences among the ensemble of possible sequences is depicted as nested sets of increasing biological relevance. Sequences having suitably well-ordered structures (prerequisite phenotype; PP) contain a subset of sequences that have the needed specific adaptations (specific phenotype; SP). Within the subset of sequences having the specific phenotype is the subset of sequences realized in the course of evolution. As defined here, self-organization defines the relative proportion of the prerequisite phenotype set with respect to the entire sequence space.

physiochemical forces that mediate interactions between the nucleotide residues inherently constrain the range and distribution of possible RNA conformations and therefore bias the range and distribution of the variation upon which selection acts. However, since the main focus of RNA structural biology has been on evolved instances of RNA, it is impossible to identify general biophysical constraints that either limit or facilitate the origin and diversification of functional RNA structures. For example, the purine-rich composition of loop regions of disparate RNAs demonstrates that the newly discovered adenosine platform structures of the P4-P6 domain of the *Tetrahymena thermophila* intron may be ubiquitous (Cate et al. 1996; Schultes et al. 1997). Cate et al. have speculated that adenosine platforms arose early in evolutionary history, presumably as an efficient mechanism for building complicated RNA architectures. Expanding the scope of comparative analysis to include random sequences generalizes traditional adaptationist hypotheses to include intrinsic constraints as a possible factor in explaining ubiquitous biological structures (Gould and Lewontin 1979). Hence, Cate and co-workers' hypothesis can be more aptly stated: Are adenosine platforms common or rare features of random RNA heteropolymers? The ubiquity of the adenosine platform in evolved RNA may be due simply to its ubiquity among random RNA polymers [these structures would be, even if deleterious, unavoidable by evolution [Kauffman 1993]. Only if adenosine platforms prove rare among random sequences does their occurrence among disparate RNA functional classes suggest an adaptive capacity in forming functional structures. Resolving these possibilities is fundamentally important to understanding the RNA folding problem and the origin of structure and function in RNA.

From a practical standpoint, the specific/prerequisite phenotype dichotomy distinguishes between order (prerequisite for function) and design (specific to function) in molecular structures (Dennett 1995) and suggests a novel approach to ribozyme engineering. Though structure prediction and rational design of complex RNA catalysts (i.e., engineering the specific phenotype) are largely intractable, combinatorial and *in vitro* evolution approaches to molecular design have had remarkable success (Jeager 1997). In principle, statistical patterns in the primary structure, e.g., base composition (Schultes et al. 1999) or sequence length (Sabeti et al. 1997), that are correlated with minimal frustration could be applied to the design of combinatorial libraries with initial populations being biased toward sequences that, though random with respect to the specific phenotype, tend to be minimally frustrated (Rejito and Verkhivker 1996; Sabeti et al. 1997). Focusing the limited diversity of the initial random-sequence pool among sequences having the prerequisite phenotype would be particularly useful when the search space is large compared to the size of the

population and when adequate solutions are exceedingly rare. The intelligent design of biased libraries represents a middle ground between rational and combinatorial approaches to RNA engineering (Wedel 1996).

*Acknowledgments.* We thank D. Bartel, J. Brown, C. Marshall, P. Unrau, B. Weber, L. Wu, and J.W. Schopf 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 the support of the Office of Naval Research acting in cooperation with the Defense Advanced Research Project Agency.

## References

- Bartel DP, Szostak JW (1993) Isolation of new ribozymes from a large pool of random sequences. *Science* 261:1411–1418
- Brown JW, Haas ES, Pace NR (1993) Characterization of ribonuclease P RNAs from thermophilic bacteria. *Nucleic Acids Res* 21:671–679
- Bryngelson JD, Onuchic JN, Socci ND, Wolynes PG (1995) Funnels, pathways, and the energy landscape of protein folding: A synthesis. *Proteins* 21:167–195
- Cate JH, Gooding AR, Podell E, et al. (1996) RNA tertiary structure mediation by adenosine platforms. *Science* 273:1696–1699
- Dennett D (1995) Darwin's dangerous idea. Simon and Schuster, New York, pp 64–65, 104–107
- Dill KA, Chan HS (1997) From Levinthal to pathways to funnels. *Nat Struct Biol* 4:10–19
- Draper DE (1996) Strategies for RNA folding. *Trends Biochem Sci* 21:165–169
- Ekland EH, Bartel DP (1995) The secondary structure and sequence optimization of an RNA ligase ribozyme. *Nucleic Acids Res* 23:3231–3238
- Ekland EH, Szostak JW, Bartel DP (1995) Structurally complex and highly active RNA ligases derived from random RNA sequences. *Science* 269:364–370
- Fontana W, Schuster P (1998) Continuity in evolution: On the nature of transitions. *Science* 280:1451–1455
- Fontana W, Konings DA, Stadler PF, Schuster P (1993) Statistics of RNA secondary structures. *Biopolymers* 33:1389–1404
- Gould SJ (1991) The disparity of the Burgess Shale arthropod fauna and the limits of cladistic analysis: Why we must strive to quantify morphospace. *Paleobiology* 17:411–423
- Gould SJ, Lewontin RC (1979) The spandrels of San Marco and the Panglossian paradigm: A critique of the adaptationist programme. *Proc R Soc Lond B* 205:581–598
- Herschlag D (1995) RNA chaperones and the RNA folding problem. *J Biol Chem* 270:20871–20874
- Higgs PG (1993) RNA secondary structure: A comparison of real and random sequences. *J Phys I France* 3:43–59
- Higgs PG (1995) Thermodynamic properties of transfer-RNA—a computational study. *J Chem Soc Faraday T* 91:2531–2540
- Hofacker IL, Fontana W, Stadler PF, Bonhoeffer LS, Tacker M, Schuster P (1994) Fast folding and comparison of RNA secondary structure. *Monatshfte Chem* 125:167–188
- Huynen MA, Stadler PF, Fontana W (1996) Smoothness within ruggedness: The role of neutrality in adaptation. *Proc Natl Acad Sci USA* 93:397–401
- Huynen MA, Guttel R, Konings DAM (1997) Assessing the reliability of RNA folding using statistical mechanics. *J Mol Biol* 267:1104–1112
- Jeager L (1997) The new world of ribozymes. *Curr Opin Biochem* 7:324–335

- Kauffman SA (1993) *Origins of order*. Oxford University Press, New York, pp 22–25
- Knuth D (1973) *The art of computer programming*, Vol 3. Addison-Wesley, Reading, MA, p 237
- Lato SM, Boles AR, Ellington AD (1995) In vitro selection of RNA lectins: Using combinatorial chemistry to interpret ribozyme evolution. *Curr Biol* 2:291–303
- Li H, Helling R, Tang C, Wingreen N (1996) Emergence of preferred structures in a simple model of protein folding. *Science* 273:666–669
- Raff R (1996) *The shape of life*. University of Chicago Press, Chicago, pp 292–320
- Rejto PA, Verkhivker GM (1996) Unraveling principles of lead discovery: From unfrustrated energy landscapes to novel molecular anchors. *Proc Natl Acad Sci USA* 93:8945–8950
- Sabeti PC, Unrau PJ, Bartel DP (1997) Accessing rare activities from random RNA sequences: The importance of the length of molecules in the starting pool. *Chem Biol* 4:767–777
- Saenger W (1984) *Principles of nucleic acid structure*. Springer-Verlag, New York, pp 143–146
- Schultes E, Hraber PT, LaBean TH (1997) Global similarities in nucleotide base composition among disparate functional classes of single-stranded RNA imply adaptive evolutionary convergence. *RNA* 3:792–806
- Schultes E, Hraber PT, LaBean TH (1999) A parametrization of RNA sequence space. *Complexity* 4:61–71
- Seilacher A (1991) Self-organizing mechanisms in morphogenesis and evolution. In: Schmidt-Kittler N, Vogel K (eds) *Constructional morphology and evolution*. Springer-Verlag, Berlin, pp 251–271
- Shakhnovich EI, Gutin AM (1990) Implications of thermodynamics of protein folding for evolution of primary sequences. *Nature* 346:773–775
- Thompson DW (1917) *On growth and form*. Cambridge University Press, Cambridge
- Unaru PJ, Bartel DP (1998) RNA-catalyzed nucleotide synthesis. *Nature* 395:260–263
- Wedel AB (1996) Fishing the best pool for novel ribozymes. *Trends Biotech* 14:459–465
- White SH (1994) Global statistics of protein sequences: Implications for the origin, evolution, and prediction of structure. *Annu Rev Biophys Biomol Struct* 23:407–439
- Zuker M, Steigler P (1981) Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res* 9:133–149