Published online October 19, 2005

5924–5935 *Nucleic Acids Research, 2005, Vol. 33, No. 18*
doi:10.1093/nar/gki886

Abundance of correctly folded RNA motifs in sequence space, calculated on computational grids

Rob Knight[1,*], Hans De Sterck[2], Rob Markel[3], Sandra Smit, Alexander Oshmyansky[4] and Michael Yarus[2]

[1]Department of Chemistry and Biochemistry and [2]Department of Molecular, Cellular and Development Biology, University of Colorado, Boulder, CO 80309-0215, USA, [3]Scientific Computing Division, National Center for Atmospheric Research, Boulder, CO 80309, USA and [4] School of Medicine, Duke University, Durham, NC 27710, USA

Received September 16, 2005; Accepted September 20, 2005

ABSTRACT

Although functional RNA molecules are known to be biased in overall composition, the effects of background composition on the probability of finding a particular active site by chance has received little attention. The probability of finding a particular motif has important implications both for understanding the distribution of functional RNAs in ancient and modern organisms with varying genome compositions and for tuning SELEX pools to optimize the chance of finding specific functions. Here we develop a new method for calculating the probability of finding a modular motif containing base-paired regions, and use a computational grid to fold several hundred million random RNA sequences containing the core elements of the isoleucine aptamer and the hammerhead ribozyme to estimate the probability that a sequence containing these structural elements will fold correctly when isolated from background sequences of different compositions. We find that the two motifs are most likely to be found in distinct regions of compositional space, and that the regions of greatest abundance are influenced by the probability of finding the conserved bases, finding the flanking helices, and folding, in that order of importance. Additionally, we can refine our estimates of the number of random sequences required for a 50% probability of finding an example of each site in unbiased random pools of length 100 to 4.1×10^9 for the isoleucine aptamer and 1.6×10^{10} for the hammerhead ribozyme. These figures are consistent with the facile recovery of these motifs from SELEX experiments.

INTRODUCTION

The abundance of sequences with particular RNA functions in random-sequence pools has important implications for the RNA World hypothesis (1), for SELEX experiments (2–4) that identify new functions from large pools of random RNA molecules, and for phylogenetic studies of highly diverged sequences that perform the same function. If sequences that can catalyze a particular reaction are especially common, the idea that the tiny amounts of RNA that would be produced by prebiotic synthesis could produce an RNA metabolism becomes more plausible. Additionally, we would expect that such sequences could continue to be evolved readily, whether in the laboratory or within modern genomes. In contrast, if sequences with particular functions are especially rare, then even rather dissimilar sequences that share a core catalytic motif would be expected to have arisen by divergent evolution from a single ancestor rather than by independent discovery of the same solution to a particular catalytic problem.

In previous work (5,6), we demonstrated that the modularity of active RNA motifs (i.e. ribozymes that catalyze specific reactions, and aptamers that bind specific targets) greatly increases their abundance in random sequences (7). Specifically, breaking up a motif into short functional modules separated by non-functional spacer greatly increases the number of ways a long sequence could contain examples of all the short pieces necessary for function, i.e. of all the modules. We demonstrated that the essential sequences for several RNA activities, such as the minimal isoleucine-binding motif (8) and the hammerhead motif (9), could be found in under 1000 zeptomoles (623 000 molecules) of RNA, raising the possibility of a zeptomole RNA world. This number is important because it is comparable to the number of RNA molecules found in a single modern cell (6).

The ability of an RNA molecule to fold into a particular secondary structure is critical for its function. One limitation of our previous work is that it only considered the core

*To whom correspondence should be addressed. Tel: +1 303 492 1984; Fax: 303 492 7744; Email: rob@spot.colorado.edu

sequence modules (defined here as only the regions where specific nucleotides are required for function), and ignored the secondary structure requirements that hold those modules in the precise juxtaposition needed for activity. In the current paper, we provide a closed-form solution to the 'module problem without pairing' that we solved algorithmically in the previous paper (6), and demonstrate the applicability of a heuristic method for solving the module problem in cases where flanking helices must exist but the precise sequences of those helices are unimportant. Additionally, we use the Zuker folding algorithm (10) as implemented in the Vienna RNA folding package (11) to assess the probability that random sequences containing the core sequence requirements for the isoleucine-binding and hammerhead motifs will also fold into the correct structure. This work assumes that correct sequence and structure are necessary for function, although the descriptions of the motifs are unlikely to be complete and the folding algorithm is imperfect. Consequently, we are using correct folding as a proxy for function, although experimentally testing the fraction of correctly folded molecules that have the predicted activity is an important future research direction.

An additional question is whether different RNA activities are most likely to be found in random-sequence pools with particular nucleotide compositions. SELEX is typically performed with chemically synthesized RNA pools where the frequencies of the 4 nt are equal, but perhaps biasing the starting pool in one direction or another would increase the probability of finding particular binding or catalytic activities. In particular, active RNA molecules are often biased towards purines (12), perhaps suggesting that a purine-rich pool would increase the abundance of functional molecules. The ability to take the background nucleotide frequency into account both for calculating the sequence probability and for folding is also important for understanding the evolution of biological RNAs such as riboswitches (13). Because the overall genome composition of organisms varies systematically over a vast range (14–16), genome composition may play a role in determining which RNA activities evolve in which species.

Thus, in this paper, we address the following questions:

(i) How common are functional RNA sequences and structures?
(ii) What is the effect of length on abundance and folding?
(iii) What is the effect of composition on abundance and folding?

Including the structural requirements as well as the sequence requirements allows more precise predictions of the number of random sequences needed to find an example of a particular active site. As in our previous work on this topic, we focus on the isoleucine and hammerhead motifs, which have been well characterized and reselected in many independent SELEX experiments (8,9,17).

MATERIALS AND METHODS

In this section, we describe the Materials and Methods we use to calculate the abundance of correctly folded RNA motifs in random sequence-pools. See Figure 1 for an overview of the process.

First we give a closed-form solution for P(sequence), the probability of finding a specific RNA motif in a random sequence, in cases where the motif does not contain random bases that must pair with each other. An efficient algorithm for calculating this probability was given in previous work (6), but we did not provide a closed-form formula there. Second, we characterize the accuracy of the Poisson approximation for estimating P(sequence) in cases where the motif contains random base pairs, which our previous calculation cannot handle. Although we previously showed that this approximation performs poorly in cases where the modules are extremely unevenly divided (6), we show here that the approximation is valid for motifs that have modularity similar to those isolated in SELEX. This new calculation takes into account the base pairs both within and between modules. We find that it is important to take base-pairing into account when calculating P(sequence), because the resulting values for P(sequence) are several orders of magnitude smaller than when base-pairing is neglected in the calculation. Third, we explain how we use computational folding to calculate P(structure|sequence), the probability that a random molecule that satisfies the sequence requirements also folds into the correct structure required for chemical function. We calculate P(structure|sequence) by computationally predicting the structures of a large sample of partially random sequences that satisfy the sequence requirements, including matching pairs of bases when the motif requires base-pairing. Finally, we show how to obtain P(structure&sequence), the probability of obtaining both the correct sequence and the correct structure in a completely random sequence, by multiplying the two probabilities calculated previously.

For the results reported in this paper, we folded ~100 million short RNA sequences using a heterogeneous computational grid composed of parallel clusters and fast workstations. This grid computing used TaskSpaces, a software framework for grid computing that we previously developed (18,19).

Closed-form solution for the RNA motif sequence probability in the absence of base-pairing

A closed-form solution for estimating the probability of finding an RNA motif without base-pairing requirements in a given RNA sequence can be derived as follows. Consider a motif with m modules, numbered from 1 to m. Denote by s_i, $i = 1, \ldots, m$ the number of spacer nucleotides preceding module i, and by s_{m+1} the number of spacer nucleotides following module m. Per definition, $s_1, s_{m+1} \geqslant 0$, and $s_i \geqslant 1$, $i = 2, \ldots, m$. The total spacer length s is then given by $s = \sum_{i=1}^{m+1} s_i$.

The probability P(sequence) of finding an RNA motif with m modules and total spacer length s in a random pool of sequences is thus approximated by

$$P(\text{sequence}) \approx \sum_{s_1=0}^{s-(m-1)} \left[(1-p_1)^{s_1} p_1 \sum_{s_2=1}^{s-s_1-(m-2)} \left[(1-p_2)^{s_2} p_2 \times \sum_{s_3=1}^{s-s_1-s_2-(m-3)} \left[(1-p_3)^{s_3} p_3 \ldots \sum_{s_m=1}^{s-s_1-\ldots-s_{m-1}} (1-p_m)^{s_m} p_m \right] \ldots \right] \right], \quad (1)$$

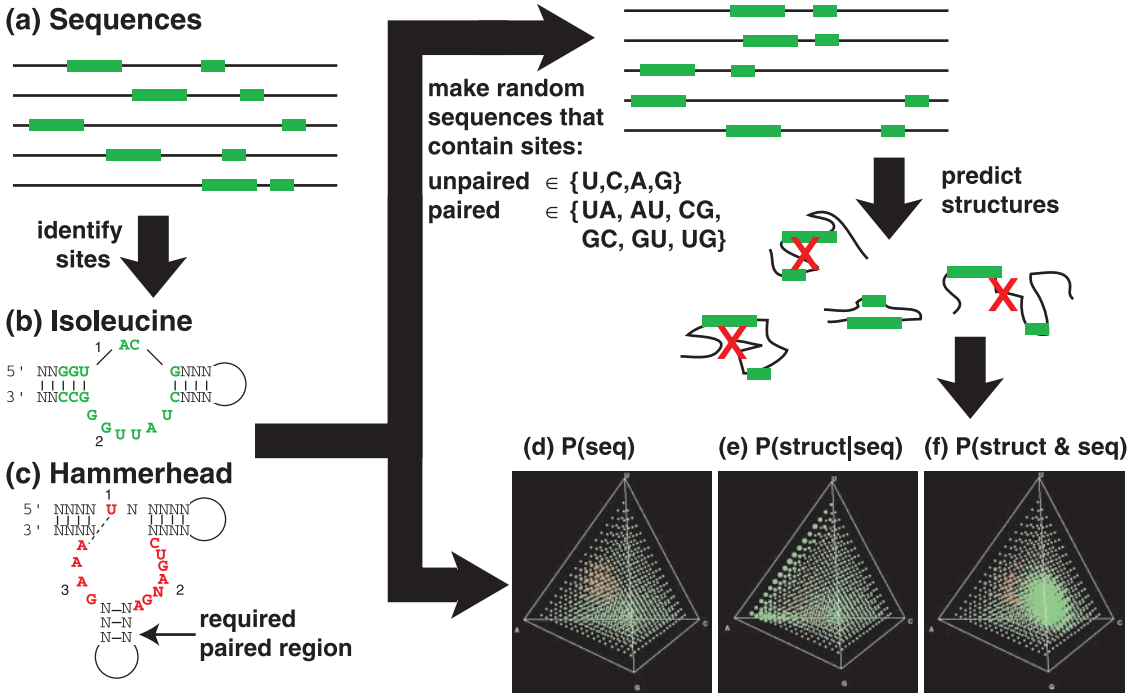

Figure 1. Procedure for determining the effects of folding and sequence composition on motif abundance. The motifs are identified by comparing sequences with the same function (**a**). The isoleucine aptamer (**b**) and the hammerhead ribozyme (**c**) both consist of modules that must have an exact sequence, and flanking helices that must base pair but need meet no other constraints. These diagrams show the exact sequence and structure requirements that were used in the calculations: base pairs are indicated by connecting lines. We had to allow an extra base pair between an A and U that are normally unpaired (dashed line) because the folding algorithm invariably extends the first helix to include this pair even though the tertiary structure of the hammerhead prevents it from forming in nature. We calculate *P*(sequence) (**d**) from the sequence requirements, and *P*(structure|sequence) (**e**) by constructing large samples of random sequences that contain the motif and computationally predicting their structures. The overall probability of finding a correctly folded sequence (**f**) is obtained by multiplying the probabilities from (d) and (e). See Figure 3 for a description of the tetrahedral simplex diagrams.

where p_i, $i = 2, \ldots, m$, are the probabilities, for the individual modules, that a random string of the module length matches the specified module.

This formula can be derived by summing the probabilities of all possible individual placements of modules

$$P_{\text{ind}} = (1-p_1)^{s_1} p_1 (1-p_2)^{s_2} p_2 \ldots (1-p_{m-1})^{s_{m-1}} p_{m-1} (1-p_m)^{s_m} p_m, \quad \mathbf{2}$$

over all possible values of the s_i, $i = 1, \ldots, m$. The individual placement probabilities can simply be added because the associated events are mutually exclusive, as the $(1-p_i)^{s_i} p_i$ factors require both the presence of modules at certain positions and their absence earlier in the sequence. Note that the formula for *P*(sequence) is not exact because the probabilities for finding the consecutive individual modules in P_{ind} are not strictly independent. However, computational experiments show that the approximation is very good for all relevant motifs and molecule lengths (6).

Heuristic methods for approximating the RNA motif sequence probability problem with pairing

The solution in Equation 1 holds for cases where the modules vary independently. However, many motifs are held together by base-paired regions that have no sequence requirements beyond those used to meet the structural requirements. This correlation between the sequences in parts of different modules poses a problem, because the probability of finding an overall match to the motif depends on which sequence matching a module's requirements was previously found. Specifically, the probability P_j of finding a module $j > i$ depends on the sequence that matches module i: this effect can be considerable when nucleotide compositions are biased.

We addressed this problem using the Poisson approximation, which assumes that the matches to each module are rare and ignores overlap. The Poisson distribution can be used to calculate the probability of obtaining a given number of matches in a sequence, once the mean number of matches per sequence is specified. In this case, we are interested in the probability that we found at least in one match. This is the complement of the probability that we found no matches, given by $Pr(X = 0) = e^{-\lambda}$, where λ is the mean number of matches per sequence. The λ is the product of p, the probability of finding a match in a single trial, and N, the number of trials in which a set of modules can be placed within a longer sequence. N is given by the formula $N = (s + 1)!/(s + 1 - m!)(m!)$, where s is the number of bases of spacer and m is the number of modules (6). The overall probability of finding at least one match in a sequence is thus $Pr(X > 0) = 1 - e^{-Np}$.

To calculate p for a specified sequence composition, we take the product of the probabilities of finding each required base and each required base pair in the motif. For example, in a random sequence containing 40% U, 20% C, 30% A, and 10% G, the probability of obtaining a particular conserved nucleotide is the frequency of the corresponding base,

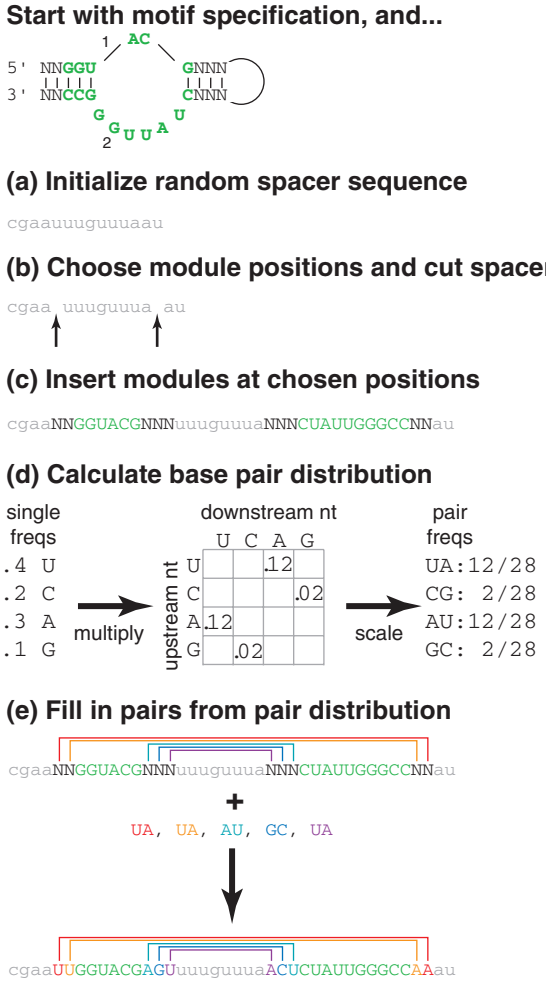

Figure 2. Procedure for making random sequences. Start with a description of a motif, including any regions whose sequence is specified and any base pairing. (**a**) The amount of spacer is calculated, and initialized with random sequence of the correct composition. (**b**) Positions for each module are chosen, and the spacer is cut at these locations. (**c**) The modules, including the specified regions of the sequence and placeholders for the paired but unspecified regions, are inserted between the pieces of spacer. (**d**) The probability of each type of base pair is calculated as the product of the frequencies of the two individual bases (the example shown is for $P(\mathrm{U}) = 0.4$, $P(\mathrm{C}) = 0.2$, $P(\mathrm{A}) = 0.3$, $P(\mathrm{G}) = 0.1$), divided by the sum of the frequencies for all valid base pairs (in the example, just the four Watson-Crick base pairs AU, UA, GC, CG). (**e**) Pairs are sampled from the pair distribution, and added to the sequence at the specified positions.

e.g. 20% for a conserved C, or 10% for a conserved G. The probability of obtaining a particular pair is the product of the probabilities of the individual bases in that pair, e.g. $0.2 \times 0.1 = 0.02$ for a GC base pair (see Figure 2d). For positions that must be paired but have no other sequence constraints, the probability of obtaining any arbitrary pair is the sum of the probabilities of obtaining each of the four possible pairs for Watson–Crick pairing, or the six possible pairs for Watson–Crick plus wobble.

We tried several additional heuristics, including a modification of our algorithm for estimating the probability of unpaired modules, but none gave results better than the Poisson over the relevant range of lengths and compositions (data not shown).

Making partially random sequences that meet structural requirements

We can calculate the probability of meeting all the sequence requirements for a set of modules using the Poisson approximation given above. Consequently, to calculate the probability of getting a correctly folded sequence by chance, we only need to multiply the probability of meeting the sequence requirements by the probability of folding into the correct structure given that the sequence requirements are met. Since the probability of meeting the sequence requirements is typically low (often $<10^{-8}$), reducing the search space to only those sequences that already met the sequence requirements (including the sequence elements required to support the necessary pairing) makes the calculations far more feasible.

We can produce a partially random sequence that meets all the sequence requirements for a motif as follows. First, randomly choose the positions within the sequence where the modules will occur. Second, fill in the spacer by choosing a random base at each position according to the individual frequencies of the 4 nt. Third, fill in any unpaired regions of each module that are partially specified by a degenerate symbol by choosing at each position a random base that could match the specification. For example, R means either A or G, so if the background composition is 20% A, 60% G, 15% C and 5% U then the resulting base will be A with $20/(20 + 60) = 25\%$ probability and G with $60/(20 + 60) = 75\%$ probability. No degenerate bases occur in the motifs examined here, but they can be important in other motifs. Fourth, calculate the distribution of valid base pairs from the individual base frequencies. This pair distribution is obtained by calculating all possible pair probabilities and dividing each pair probability by the sum of the probabilities for those pairs that are considered valid: either the four Watson–Crick pairs, or these four plus GU and UG if wobble is included. Finally, fill in the paired positions by sampling pairs from this distribution, using the first member of the pair for the upstream position and the second member of the pair for the corresponding downstream position. Figure 2 describes this process.

We focused on two motifs: the isoleucine aptamer motif (8,17), which is the simplest RNA motif that can bind the amino acid isoleucine, and the hammerhead ribozyme motif, an example of a self-cleaving motif that has been isolated both from organisms and from SELEX (9). The minimal sequence and structural requirements for these motifs are shown in Figure 1.

Calculating *P*(structure|sequence) using computational folding on heterogeneous computational grids

This study required us to estimate structures for ~100 million short RNA sequences (50–100 nt length). For a given composition and molecule length, we typically used a sample size of 10 000 partially random molecules to estimate *P*(structure|sequence). Stepping through composition space in 5% intervals (969 different points) and variations in molecule length and problem definition led to a total of ~100 million computational foldings. This constitutes a computational problem of moderately large size, which would require weeks to months on a single fast workstation. We decided to use a grid computing approach, mainly for flexibility, portability and scalability. In grid computing, CPU cycles provided by computers of

various sizes and types are treated as an exchangeable commodity, accessible through a standard interface, in close analogy with power grids. A primary advantage of grid computing is inherent scalability: in a grid system the grid operator can add 'compute farms' to provide extra capacity on demand. Stand-alone workstations, clusters, and parallel supercomputers can be combined seamlessly depending on the size of the problem and the desired turn-around time.

Several grid computing environments have been developed in recent years. The results presented below were obtained using the TaskSpaces software framework for grid computing (18,19). TaskSpaces is a lightweight grid computing framework for scientific computing implemented in Java. Application code need not be installed and maintained on worker machines, because it is downloaded from a central server when task objects arrive at the workers. Installing and executing a Java bytecode executable of size <2 kB allows any worker host to participate in the grid. Thus, installation and maintenance of TaskSpaces is easy.

We used the RNAfold program in the Vienna RNA folding package (11), which is written in C, for folding individual sequences. A sequence was considered to have folded correctly if the minimum free energy structure at 37°C using the default energy parameters includes all of the pairs required by the motif specification, and includes no base pairs that involved positions required to be unpaired in the motif specification. The RNAfold executable is called by the Java application on each worker node as needed. Non-Java executables must be compiled in advance for each worker architecture, and can be downloaded from the code server by the workers upon first use. Thus, although reliance on code written in other languages increases the effort required for cross-platform operation, it is still feasible.

Our results were obtained on a grid composed of the NCSA IA32 Linux Platinum Supercluster and various P4 Linux workstations at CU Boulder. The Platinum machine features 968 P3 compute processors (1 GHz). For some smaller problems, only the local workstations were used, while for larger problems the local workstations were combined with up to 200 Platinum processors concurrently. The total computing time used for this project so far, including extensive initial runs for exploring the problem and determining the right approach and questions to be answered, amounts to approximately 10 000 Platinum processor hours.

Center of mass calculations

To compare any two distributions, we scaled the probabilities of the points in each of the distributions linearly so that the most probable point had a value of 1. We calculated the centers of mass of the two distributions, and calculated the Euclidean distance between the centers of mass. For each possible location in the simplex, there are thus two values, one from each distribution. We randomly reassigned the values at each location among series, so that each new series had the same number of values in the same locations as the original series but the associations between values and series were randomized. We calculated the distance between the centers of mass in the new, randomized distributions, and compared this distance to the distance obtained for the real data. By repeating this procedure many times, we could find the probability that the distance between two randomized distributions would exceed the distance between the two actual distributions. If this probability was small, we rejected the hypothesis that the two actual distributions were the same.

Display of compositional data

We needed to display results that covered the full volume of compositional space. Although there are 4 nt, their frequencies are constrained by a sum, leading to only three degrees of freedom in composition. Consequently, it is possible to plot the compositions of the four bases in three dimensions without loss of information, using the simplex method (12). We constructed the simplex by using the three orthogonal axes of pairwise nucleotide composition: AU versus GC (the weak–strong axis, also called Chargaff's axis), AG versus UC (the purine–pyrimidine axis), and AC versus GU (the amino–keto axis) (Figure 3a). These axes define a cube in which four of the vertices correspond to pure homopolymers of A, C, G and U; however, the remaining four vertices cannot correspond to the composition of any actual sequence, because the constraints would be incompatible (Figure 3b). Connecting the vertices corresponding to the homopolymers gives a tetrahedron within which all actual sequence compositions lie. Compositions that have more of a particular nucleotide lie closer to the vertex for that nucleotide. We plotted our data on a 5% grid of composition space, with a point for each composition that is an even multiple of 5% in each of A, C, G and U. In general, the volume of each point on these diagrams is proportional to a probability we have calculated at that composition.

Implementation details

We implemented all randomization and structure assessment code three times, twice independently in Python and once in Java, to ensure consistent results across platforms and random number generators. Additional code to test the different heuristic methods for estimating the match probabilities and for output was written in Python. Results were displayed using the MAGE visualization package (20). Source code is available from the authors on request.

RESULTS

Accuracy of the Poisson approximation

We tested the validity of the Poisson approximation by Monte Carlo simulation, using a two-module example in which each module consisted of a single unpaired region and a single helix. We varied the length of the spacer, the length of the unpaired region, and the length of the helix. The expected number of matches was compared with the actual number of matches obtained by creating completely random sequences, searching the sequences for the constant motifs, and counting the number of sequences in which the helix could form through complementary base pairing.

Figure 4 shows representative results for varying the helix length with a fixed constant region of 3 nt and fixed spacer of 20 nt (Figure 4a), varying the constant region with a fixed helix length of 3 bp and fixed spacer of 20 nt (Figure 4b),

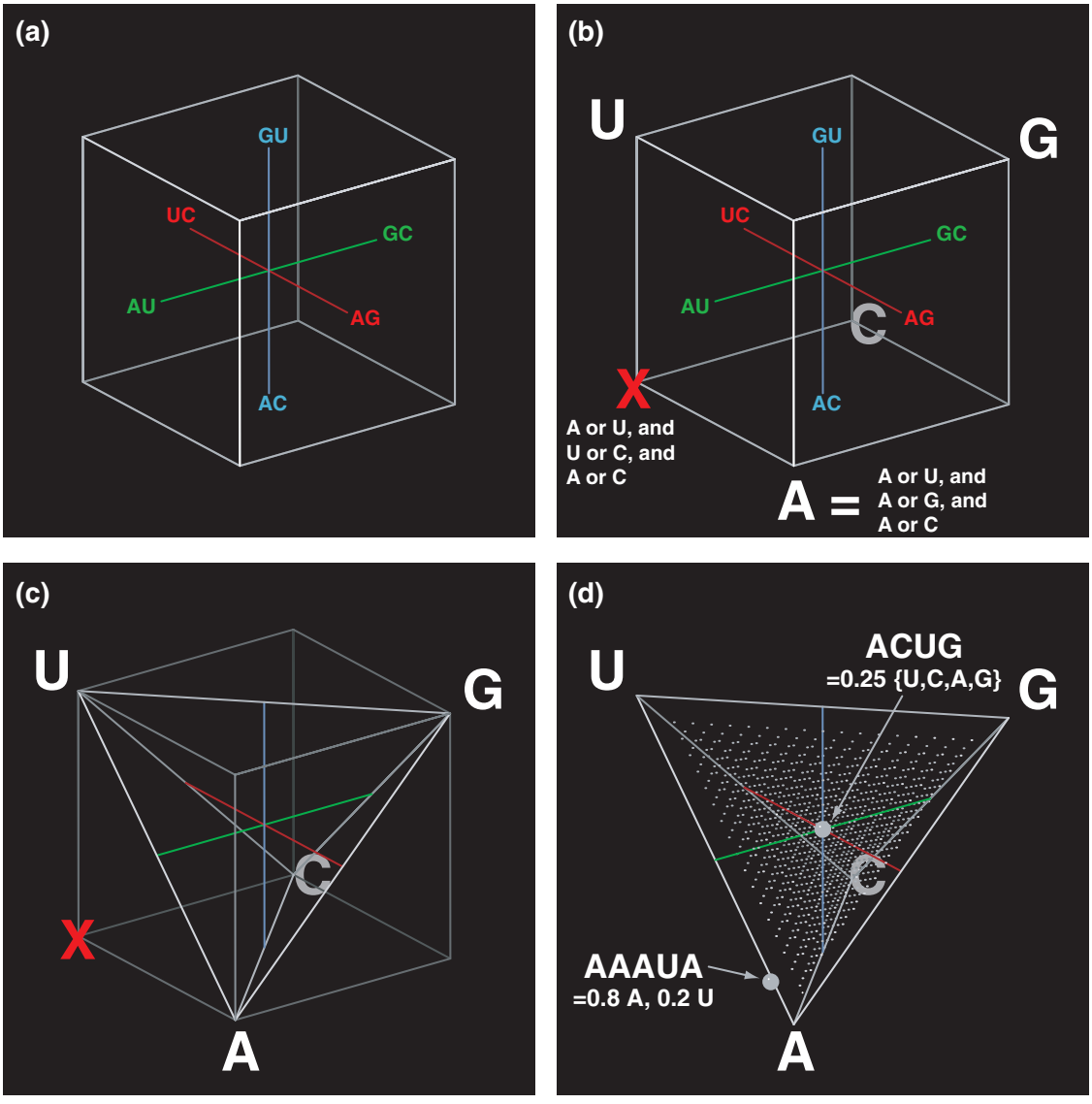

Figure 3. Plotting compositional results in the RNA Simplex (12). RNA composition can be defined in terms of three orthogonal axes (**a**) the weak-strong axis (or Chargaff's axis), GC versus AU; the purine–pyrimidine axis, AG versus CU; and the amino–keto axis, AC versus GU. These three axes define a cube, with composition ranging from 0 to 1. For example, a sequence composed entirely of G and C would lie at 1.0 on Chargaff's axis, a sequence composed entirely of A and U would lie at 0.0, and a sequence with equal amounts of G + C and A + U would lie at 0.5. These three axes define the points at which pure homopolymers of each nucleotide lie (**b**) for e.g., a sequence that is 100% A is 100% A + U and 100% A + G and 100% A + C, so must lie at the corresponding extreme on each axis. In contrast, some points within the cube cannot correspond to actual sequences: the X marks a sequence that would have to be 100% A + U, and 100% A + C, and 100% U + C, which is impossible. Connecting the vertices corresponding to the homopolymers gives a tetrahedron (**c**), within which all actual sequence compositions must lie. The composition of specific sequences (**d**) can be calculated by locating the coordinates corresponding to the sum of G + C, G + A, and G + U in that sequence. Most of our calculations use a 5% composition grid, in which all compositions that are an even multiple of 5% in each of U, C, A and G are present. The volume of the point is proportional to the probability of folding or abundance, depending on the calculation.

varying the spacer with the constant region fixed at three nucleotides and the helix length fixed at three base pairs (Figure 4c), and varying the nucleotide composition (Figure 4d). The Poisson approximation recaptured the observed results within sampling error in the range shown here, although the error was somewhat more pronounced when the constant regions and/or helices were short.

Probability and folding results

To test the effects of nucleotide composition on the probability of meeting the sequence requirements and the probability of correct folding, we generated 10 000 random sequences at each of the 969 possible 5% intervals of sequence composition. The total length of the sequences were 50, 100 and 150 nt, meeting the sequence requirements for each of the hammerhead and isoleucine motifs. We repeated the analysis for sequence length 50 allowing GU base pairs. Thus we folded a total of 77 520 000 sequences for this experiment. Except where otherwise indicated, results will refer to these samples of 10 000 sequences at each composition.

Figure 5 shows, for all 5% intervals of nucleotide composition in the space of possible compositions, the probability of meeting the sequence requirements in a completely random

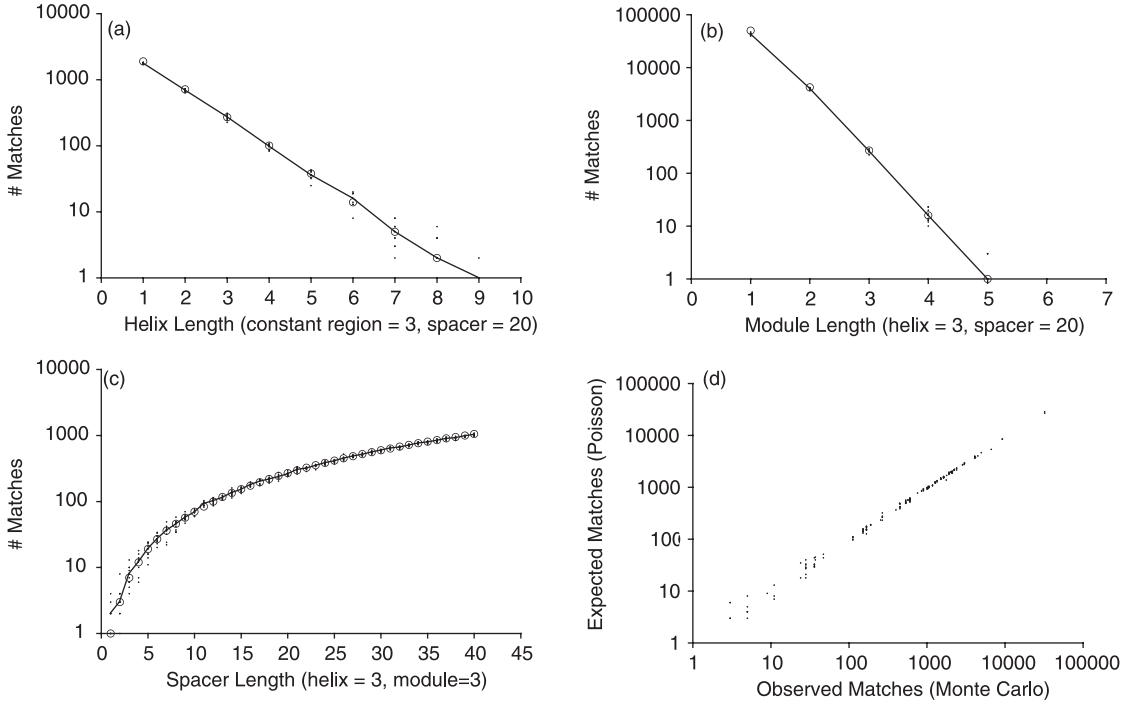

Figure 4. Performance of Poisson heuristic used to predict paired motif abundance. The *y*-axis always shows the number of matches for two paired modules in samples of 100 000 random sequences. Individual Monte Carlo simulation runs are shown as small black dots, with a line connecting the mean of each set of simulation runs where appropriate (**a–c**). The estimated numbers of matches from the Poisson heuristic are shown as blue circles. All simulations shown use two modules, and, unless otherwise indicated, equal base frequencies. (**a**) Effect of helix length (*x*-axis), ranging from 1- to 6-bp. Each module additionally contained three fully specified nucleotides, and the modules were embedded in 20 nucleotides of spacer. (**b**) Effect of module length (*x*-axis): the length of the constant part of each module was varied from 1 to 7 nt, embedded in 20 nt of spacer with a helix length of 3. (**c**) Effect of spacer length (*x*-axis): the length of the spacer was varied from 1 to 40 nt, with three fixed nucleotides in each module and a 3 bp helix connecting them. (**d**) Robustness of results to changes in nt composition: a 3 base helix and two 3 base constant regions were embedded in 50 nt of spacer in a total of 21 background compositions ranging from unbiased to 70% of the most common nucleotide and 10% of each other nucleotide. Graph shows observed count (*x*-axis) against predicted count (*y*-axis) for 210 samples using different random choices of bases for the fixed regions. These motifs, depending on the background composition, have predicted frequencies ranging from <1 in 100 000 to 32%.

sequence for the hammerhead and isoleucine motifs (red and green respectively; Figure 5a and b), the probability of correct folding in partially random sequences that already meet the sequence requirements (Figure 5c and d), and the combined probability of finding the correctly folded motif. In sequences of total length 100, including GU pairs, the probability of finding the isoleucine motif ranged from 1.44×10^{-21} to 5.71×10^{-10} with a mean of 3.62×10^{-11}, reaching a value of 1.71×10^{-10} at unbiased nucleotide frequency and a maximum at the coordinates 15% A, 25% C, 35% G, and 25% U. The probability of finding the hammerhead motif ranged from 0 to 4.58×10^{-10} with a mean of 7.37×10^{-12}, reaching a value of 3.38×10^{-11} at unbiased nucleotide frequency and a maximum at the coordinates 35% A, 10% C, 25% G, and 30% U.

Figure 6a and b shows the effect of allowing GU wobble pairs in the helices on the combined probability. Wobble pairing increases the probability of meeting the sequence requirements, but dramatically decreases the probability of correct folding (especially for the hammerhead, which has longer paired regions). These results are for total sequence length 50.

Figure 6 shows the effect of length on the overall probability, showing the overall probability of correct folding at length 50 (a and b), 100 (c) and 150 (d), all including GU pairs except (b). The location of maximum probability remains essentially unchanged, as does the center of mass for the probability, although the overall probability increases substantially as the length increases (from 4.27×10^{-12} to 8.62×10^{-10} for the hammerhead motif, and from 1.88×10^{-10} to 1.06×10^{-9} for the isoleucine motif).

Finding the points of highest motif probability

In order to refine our estimates of the maximum probability of each motif, we folded 100 samples of 10 000 sequences (a total of 1 million sequences), each 100 nt long, at each of three compositions for each motif: the composition at which abundance was maximal, the composition of the conserved nucleotides, and the unbiased composition. We then used two-sample *t*-tests on the sample of estimated probabilities at each composition to test whether these probabilities differed significantly. For the hammerhead motif, the optimal composition identified in the initial folding was 35% A, 10% C, 25% G, and 30% U. The definition of the hammerhead motif we used contained 5 conserved As, 1 C, 3 Gs, and 2 Us, giving frequencies of 45.5%, 9.1%, 27.3% and 18.2% respectively. At the optimal composition, the mean overall probability of finding the hammerhead motif was 1.02×10^{-10}, a factor of 2.5 to 6 times greater than at the conserved or unbiased compositions (1.6×10^{-11} and 4.5×10^{-11} respectively). These differences were highly significant: $P = 2.5 \times 10^{-16}$

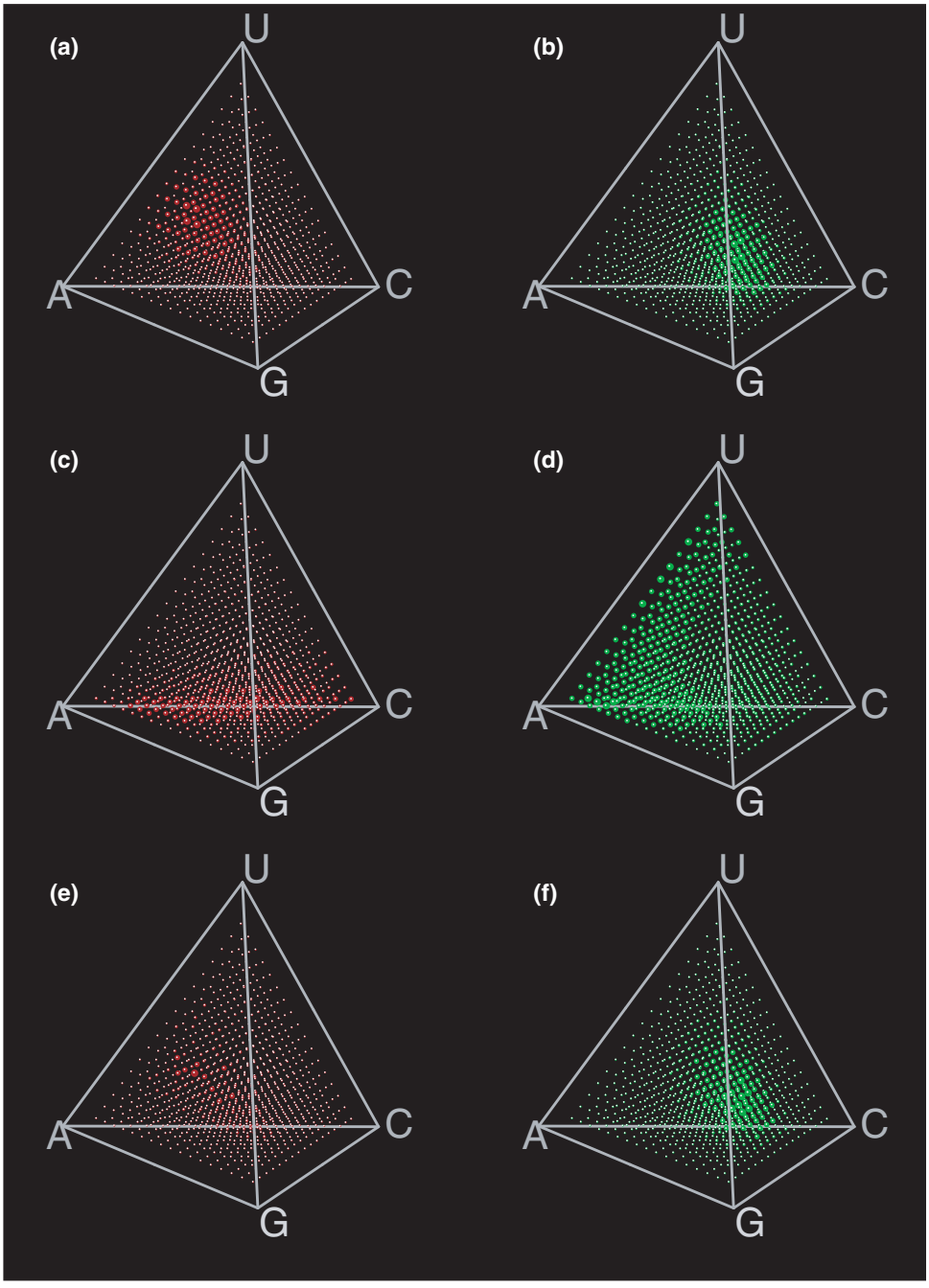

Figure 5. Folding results for the Hammerhead (red) and Isoleucine (green) motifs. Probability of finding the required sequence elements (**a** and **b**), probability of folding correctly given that the required sequence elements were present (**c** and **d**), and overall probability of having the required sequence elements and folding correctly (**e** and **f**). Volume of each sphere is proportional to the probability at each of the 969 internal 5% intervals in the space of possible compositions. Radii are scaled such that the maximum radius in each diagram is set to 0.01 composition unit. These results are for sequence length 100.

for the comparison between the optimum point and the conserved frequencies, and 3.1×10^{-9} for the comparison between the optimum point and unbiased frequencies.

For the isoleucine motif, the differences were much less pronounced. The optimal composition was 15% A, 25% C, 35% G, and 25% U. However, the motif contained 2 conserved As, 4 Cs, 6 Gs and 4 Us, giving frequencies of 12.5%, 25%, 37.5% and 25%, respectively: closer to the point we identified as the optimum than for any other grid point. At the optimal composition, the mean overall probability of finding the isoleucine motif was 5.9×10^{-10}, which was slightly less than at the conserved composition value of 6.4×10^{-10}. However, the probability of finding the isoleucine motif at unbiased composition was over 3-fold lower: 1.7×10^{-10}. These differences were also highly significant: $P = 5.4 \times 10^{-15}$ for the comparison between the optimum point and the conserved frequencies, and 7.8×10^{-112} for the comparison between the optimum point and unbiased frequencies. These results

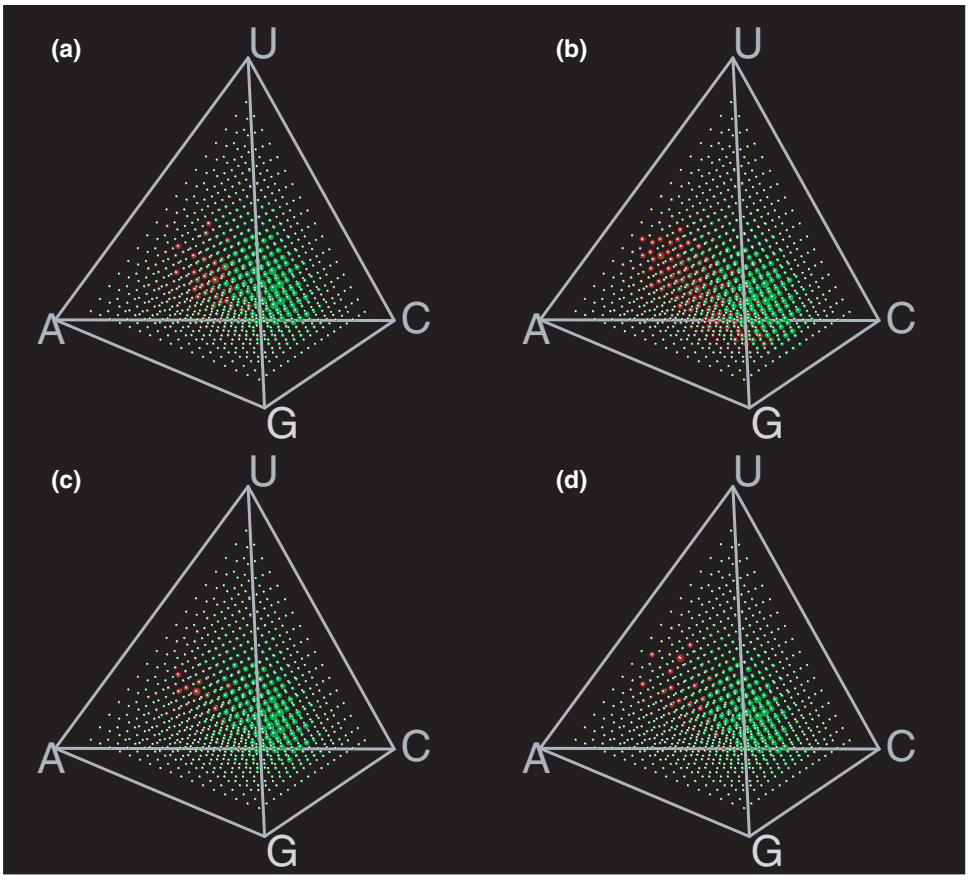

Figure 6. Folding results for the Hammerhead (red) and Isoleucine (green) motifs. Overall probability of finding the correctly folded motif in (**a**) sequences of length 50, (**b**) length 50 disallowing GU pairs, (**c**) length 100, and (**d**) length 150. Sequence length does not affect the region of composition space in which each motif is most likely to be found, although there are some variation due to sampling effects. Volume of each sphere is proportional to the probability at each of the 969 internal 5% intervals in the space of possible compositions. Radii are scaled such that the maximum radius in each diagram is set to 0.01 composition unit.

demonstrate that although the optimal composition is dominated by the necessity of finding the conserved bases in the motif, as expected, the requirement to find the flanking helices and the folding also affect the result significantly. The hammerhead motif, which requires longer flanking helices than the isoleucine motif (12 bp versus 9), shows a large difference between the optimal composition and the composition of the conserved bases.

In order to test how the probability of finding each motif varied close to its optimum point, we examined the gradient of change away from the initially determined optimum point for both motifs. Specifically, we tested for differences in overall probability at each of the 5% points neighboring the optimal point for each motif, using samples of 50 000 sequences at each neighboring point and one million sequences at each optimal point, using sequences of length 100 and allowing GU pairs. The isoleucine motif decreased in probability primarily when G was replaced by A or C, and when U was replaced by A, the differences in these directions being up to 1.9-fold greater than the differences in other directions. The gradient of change was much smaller and more even for the hammerhead motif than for the isoleucine motif, with many of the adjacent points being statistically indistinguishable from the initially identified optimum point. The largest difference for the hammerhead motif was in replacing C with G, and the smallest when replacing G with A.

To test whether the compositional grid was sufficiently fine to locate the region of maximum probability, we performed a more detailed analysis of the transect between the two best-folding points for the isoleucine aptamer using a larger sample size of 100 000 sequences per point to reduce the effects of sampling error. Figure 7 shows the folding probabilities at 40 intervals between the two best points at sequence length 50: 10% A, 30% C, 35% G, 25% U and 10% A, 30% C, 40% G, 20% U. The monotonic increase in folding efficiency between the second-best and best point suggests that the probability of correct folding is relatively smooth.

Finally, we calculated the point that maximized the probability of finding both the isoleucine and hammerhead motifs by calculating the pool size required for 99% probability of occurrence of each motif according to the method from eknig03a (21), taking the maximum of the two pool sizes for each composition. The minimum pool size over all compositions was at 25% U, 15% C, 20% A, and 40% G. This optimal point required 6.23×10^9 molecules for 99% abundance of both motifs: many orders of magnitude less RNA than is typically used in SELEX. In comparison, an unbiased pool required 2.93×10^{10} molecules for 99% abundance.

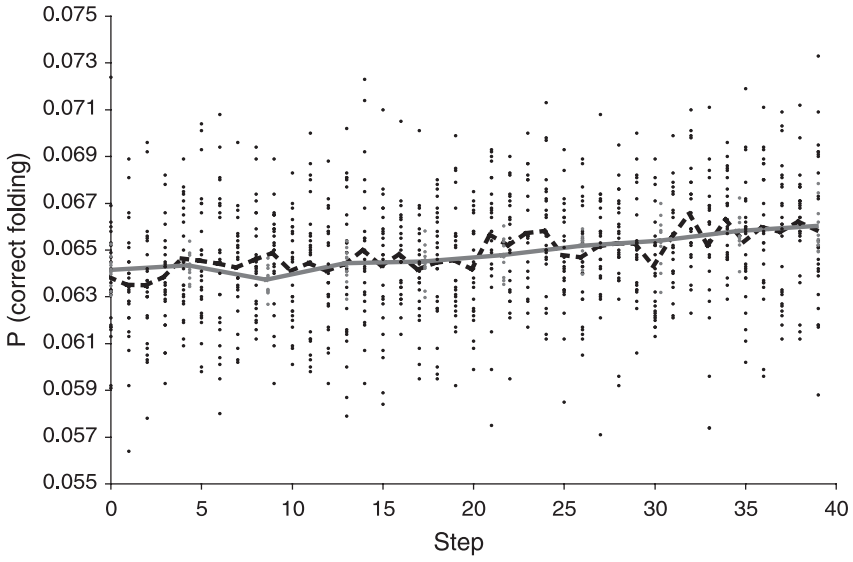

Figure 7. Fine-grained analysis of regions between the two most probable points for isoleucine aptamer folding, which were 10% A, 30% C, 35% G, 25% U and 10% A, 30% C, 40% G, 20% U using a total sequence length of 50 nt. We made 10 independent samples, each of 100 000 sequences, at each of 10 equal intervals between the two most probable points (blue dots; blue line shows the mean), and made 25 independent samples, each of 10 000 sequences, at each of 40 equal intervals between these same two points (black dots; red line shows the mean). Both series are shown at the same scale in terms of absolute composition. The lines for the means are smooth in both cases, although (as expected) the scatter is lower for the points at the larger sample size.

Table 1. Single-sequence match probabilities and pool sizes required for 50% abundance: at equal base frequencies: comparison between this analysis and our previous calculations (6)

Motif	New	Old Min	Old Mean	Old Max
Ile Probability	1.7×10^{-10}	2.1×10^{-4}	3×10^{-6}	5.0×10^{-8}
Num Seqs	4.1×10^{9}	3.3×10^{3}	2.1×10^{5}	4.5×10^{8}
HH Probability	4.5×10^{-11}	3.1×10^{-3}	6.2×10^{-11}	2.2×10^{-18}
Num Seqs	1.6×10^{10}	2.2×10^{2}	1.1×10^{10}	3.1×10^{17}

Columns: New gives the new calculations. Old Min, Old Mean, and Old Max give the previous calculations depending on which nucleotides are counted as part of the site: Min counts only completely specified nucleotides, Max counts all nucleotides that form part of the structural requirements but are otherwise unspecified as though they were completely specified and Mean counts half of the nucleotides involved in structural requirements as though they were completely specified. The large range for the Hammerhead site comes from the fact that it contains three long helices that are free to vary as long as they meet the structural requirements.

Interestingly, 35 compositions of the 969 had higher probabilities of finding both motifs than did an unbiased pool. On average, these compositions were depleted in C (average 18.1%), and enriched in A and G (average 26.6 and 30.6%). These observations are consistent with the findings that C is the easiest base to eliminate from a functional RNA molecule (22), and that functional RNA molecules are often purine-biased (12).

Together, these experiments allow us to refine our estimate of the probability of finding correctly the folded sequences considerably. Table 1 compares the estimates, including the folding, from this work with our previous estimates of the abundance of the same motifs (6).

DISCUSSION

Biases in nucleotide composition had substantial, and sometimes opposing, effects on both the probability of meeting the sequence requirements of a motif and the probability of correct folding. The locations of optimal sequence abundance and optimal folding differed from each other and from the location of highest overall probability for both of the two motifs we studied. Pairwise comparisons of the centers of mass showed that these differences were significant ($P < 0.001$ in all cases). Interestingly, the hammerhead -motif, which contains far more base pairs than the isoleucine motif, was predominantly found in AC-biased regions, in which sequences forming the long helices required for function were difficult to find because A and C do not pair with each other. It is plausible that the competition from alternative folds made this motif more difficult to recover in less biased regions where more possibilities for base-pairing exist. This observation suggests a general principle affecting motifs that require long helices or many helices for their function, which is that such motifs are most likely to be found in compositional regions where overall base-pairing is minimized.

The optimal nucleotide compositions could make a substantial difference in the overall probability of finding the motifs: at sequence length 100, 2.3-fold for the hammerhead motif and 3.5-fold for the isoleucine motif compared to unbiased frequencies, and 6.4-fold for the hammerhead motif compared to the frequencies of the conserved bases. A difference of this magnitude might be possible to detect experimentally. The hammerhead motif and the isoleucine motif had also rather different optimal compositions, suggesting that it may be possible to favor one activity or another by tuning the composition of the SELEX pool. Although the optimal composition for the isoleucine motif could have been predicted from the conserved bases at the active site, the optimal composition for the hammerhead was substantially different, primarily because of the long flanking helices required for activity. Repeating our analysis for a larger selection of motifs should provide more detailed information about

whether such discriminatory SELEX can recover particular desired functions with higher efficiency.

The fine-grained analysis (Figure 7) showed that the effects of composition on folding are smooth, and that we can consequently be confident that we have found the regions of composition space with the highest folding probabilities with sufficient accuracy. This conclusion is useful for future analyses, since the number of points in the space of possible compositions increases as the cube of the reciprocal of the interval: there are 969 possible non-edge points at 5%, but 156 849 possible points at 1%. If the effect of composition had been rugged, identification of the optimal regions could require many orders of magnitude more CPU time.

The random sequence length had little to no effect on the nucleotide compositions at which the motifs were most likely to be found, or on the folding efficiency. Thus the regions of maximum probability were within one step of each other across sequence lengths. This suggests that any general rules about regions of composition space that promote correct folding will apply to sequences of any length.

As expected (5–7), longer sequences had a large combinatorial advantage over short sequences in meeting the sequence requirements (maximum probabilities of 1.74×10^{-8}, 1.42×10^{-6}, and 7.87×10^{-6} for 50, 100, and 150 nucleotides for the hammerhead motif, and 3.46×10^{-9}, 3.20×10^{-8} and 8.94×10^{-8} for isoleucine: the probability for isoleucine aptamer changes more slowly because it has two modules instead of three for the hammerhead). However, this combinatorial advantage was offset somewhat by substantially worse folding at greater sequence lengths (maximum probabilities of 5.64×10^{-2}, 2.42×10^{-2}, and 1.08×10^{-2} for 50, 100, and 150 nt for the hammerhead motif, and 3.17×10^{-1}, 1.78×10^{-1} and 1.29×10^{-1} for isoleucine). The maximum overall probabilities for the two sites were 4.27×10^{-12}, 1.02×10^{-10}, and 8.61×10^{-10} for 50, 100, and 150 nt for the hammerhead motif, and 1.88×10^{-10}, 5.93×10^{-10} and 1.06×10^{-9} for isoleucine (note that these are not the products of the best probabilities for finding the sequence requirements and for folding, because the optima occurred at different compositions). Thus, longer sequences have a combinatorial advantage in allowing more matches to be found, although the rate of increase of this advantage decreases as the sequence length increases.

These findings are difficult to reconcile with experiments which show that certain motifs can be much more difficult to isolate from longer random regions (8,23). One possibility is that the computational folding systematically overestimates the probability of a correct fold in longer sequences; another is that other effects of sequence length, notably amplification efficiency, outweigh the effects of function at the RNA level. We plan to test these effects directly by synthesizing sequences that are computationally predicted to fold into one motif or the other. We will then use chemical and enzymatic probing to test the structural predictions around each motif, and assay the relevant catalytic and binding parameters to determine whether the molecules perform the predicted function.

The motif probabilities reported here are several orders of magnitude lower than those reported for the same motifs in previous work (6). The primary reason for this discrepancy is that we had previously excluded the contribution of paired regions whose sequence was otherwise unspecified. Incorporating these structural requirements both in the definition of the motif and also by computationally folding the sequences greatly reduced the probability of both the isoleucine and hammerhead motifs. Although the predicted frequencies are certainly compatible with the pool sizes of 10^{13}–10^{15} molecules typically used in SELEX, they fail to support the idea that an RNA world could have begun with mere zeptomoles of RNA.

CONCLUSIONS

The two motifs studied here show striking differences in the sequence compositions from which they are most likely to be recovered. These differences, although partly predictable from the different sequence requirements at the conserved sites, suggest that SELEX experiments may be tuned to prefer particular kinds of solution by the expedient step of biasing the nucleotide composition in the starting pool. When the analysis is repeated for a larger, more diverse database of motifs, we may find more general rules for optimizing SELEX outcomes. In particular, it may be possible to reduce the abundance of particular undesirable kinds of motifs, such as those conferring column affinities or enhancing amplification in transcription or PCR, thereby surviving the selection process without performing the desired function.

The inclusion of the contribution of regions that must be base-paired but are otherwise unspecified greatly affects the probability of finding both the sequence and the structure requirements for particular motifs. Consequently, we have refined our estimates of the number of random RNA molecules that must be searched to find the isoleucine motif from between 3.3×10^3 and 4.5×10^8 to about 4.1×10^9, and to find the hammerhead motif from between 2.2×10^2 and 3.1×10^{17} to about 1.6×10^{10}. These figures are for a 50% probability of finding the motif in 100 nt sequences in unbiased random-sequence pools.

These numbers are consistent with the facile recovery of these motifs from SELEX experiments, but are orders of magnitude smaller than the size of the random-sequence pools typically used in SELEX. They are inconsistent with our earlier proposal that the RNA world could have started with zeptomole or attomole pools (5,6), suggesting instead that femtomoles of RNA (tens of nanograms) would mark the threshold for evolution of aptamers and ribozymes, and therefore would be the smallest pools useful for a ribocyte. Owing to the chemical problems in synthesizing large amounts of RNA without enzymes, it has often been suggested that a simpler self-reproducing system preceded the RNA World. However, the amounts of RNA we predict to be required for function are still very small: 1.6×10^{10} 100 nt RNA molecules is about 0.8 nanograms of RNA, about the quantity of RNA found in a sample of 1.5×10^{-8}g of modern bacteria, or about 15 000 cells. Consequently, once RNA was first synthesized (perhaps for an entirely different reason), our results show that catalytic activity would soon be likely to emerge.

ACKNOWLEDGEMENTS

Some parts of this work were supported by NIH research grant GM 30881 and NASA Center for Astrobiology grant

NCC2-1052, by National Computational Science Alliance grant MCB020011 using the NCSA IA32 Linux Supercluster 'Platinum' and by NSERC Discovery and RTI grants RGPIN311947-05 and EQPEQ316028-05. We thank Erik Schultes and members of the Knight and Yarus labs for critically reading the manuscript, and two anonymous reviewers for suggesting additional analyses. Funding to pay the Open Access publication charges for this article was provided by the NIH and NASA grants cited above.

Conflict of interest statement. None declared.

REFERENCES

1. Gilbert,W. (1986) The RNA world. *Nature*, **319**, 618.
2. Tuerk,C. and Gold,L. (1990) Systematic evolution of ligands by exponential enrichment: RNA ligands to Bacteriophage T4 DNA polymerase. *Science*, **249**, 505–510.
3. Robertson,D.L. and Joyce,G.F. (1990) Selection *in vitro* of an RNA enzyme that specifically cleaves single-stranded DNA. *Nature*, **344**, 467–468.
4. Bartel,D.P. and Szostak,J.W. (1993) Isolation of new ribozymes from a large pool of random sequences. *Science*, **261**, 1411–1418.
5. Yarus,M. and Knight,R. (2000) The scope of selection. In de Pouplana,L Ribas (ed.), *The Genetic Code and the Origin of Life*. Landes Bioscience, Georgetown, TX, USA, pp. 75–91.
6. Knight,R. and Yarus,M. (2003) Finding specific RNA motifs: function in a zeptomole world? *RNA*, **9**, 218–230.
7. Sabeti,P.C., Unrau,P.J. and Bartel,D.P. (1997) Accessing rare activities from random RNA sequences: the importance of the length of molecules in the starting pool. *Chem. Biol.*, **4**, 767–774.
8. Lozupone,C., Changayil,S., Majerfeld,I. and Yarus,M. (2003) Selection of the simplest RNA that binds isoleucine. *RNA*, **9**, 1315–1322.
9. Salehi-Ashtiani,K. and Szostak,J.W. (2001) *In vitro* evolution suggests multiple origins for the hammerhead ribozyme. *Nature*, **414**, 82–84.
10. Zuker,M. and Stiegler,P. (1981) Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.*, **9**, 133–148.
11. Hofacker,I., Fontana,W., Stadler,P.F., Bonhoeffer,L.S., Tacker,M. and Schuster,P. (1994) Fast folding and comparison of RNA secondary structures. *Monatshefte Fur. Chemie*, **125**, 167–188.
12. Schultes,E., Hraber,P.T. and LaBean,T.H. (1997) Global similarities in nucleotide base composition among disparate functional classes of single-stranded RNA imply adaptive evolutionary convergence. *RNA*, **3**, 792–806.
13. Winkler,W., Nahvi,A. and Breaker,R.R. (2002) Thiamine derivatives bind messenger RNAs directly to regulate bacterial gene expression. *Nature*, **419**, 952–956.
14. Sueoka,N. (1962) On the genetic basis of variation and heterogeneity of DNA base composition. *Proc. Natl Acad. Sci. USA*, **48**, 582–592.
15. Muto,A. and Osawa,S. (1987) The guanine and cytosine content of genomic DNA and bacterial evolution. *Proc. Natl Acad. Sci. USA*, **84**, 166–169.
16. Knight,R.D., Freeland,S.J. and Landweber,L.F. (2001) A simple model based on mutation and selection explains trends in codon and amino-acid usage and gc composition within and across genomes. *Genome Biol*, **2**, RESEARCH0010.
17. Majerfeld,I. and Yarus,M. (1998) Isoleucine:RNA sites with associated coding sequences. *RNA*, **4**, 471–478.
18. De Sterck,H., Markel,R.S., Pohl,T. and Rüde,U. (2003) A lightweight Java TaskSpaces framework for scientific computing on computational grids. *Proceedings of the ACM Symposium on Applied Computing, Track on Parallel and Distributed Systems and Networking*, Melbourne, FL, 9-12 March 2003, 1024–1030.
19. De Sterck,H., Markel,R.S. and Knight,R. (2005) TaskSpaces: a software framework for parallel bioinformatics on computational grids. In Zomaya,A. (ed.), *Parallel Computing in Bioinformatics and Computational Biology*. John Wiley and Sons, Georgetown, TX, USA, pp. 1024–1030.
20. Richardson,D.C. and Richardson,J.S. (1992) The kinemage: a tool for scientific communication. *Protein Sci.*, **1**, 3–9.
21. Knight,R. and Yarus,M. (2003) Analyzing partially randomized nucleic acid pools: straight dope on doping. *Nucleic Acids Res.*, **31**, e30.
22. Rogers,J. and Joyce,G.F. (1999) A ribozyme that lacks cytidine. *Nature*, **402**, 323–325.
23. Huang,F., Bugg,C.W. and Yarus,M. (2000) RNA-catalyzed CoA, NAD, and FAD synthesis from phosphopantetheine, NMN, and FMN. *Biochemistry*, **39**, 15548–15555.