[Frontiers in Bioscience 13, 6060-6071, May 1, 2008]

Information, probability, and the abundance of the simplest RNA active sites

Ryan Kennedy[1,2], Manuel E. Lladser[2], Michael Yarus[3], Rob Knight[4]

[1]Department of Computer Science, University of Colorado at Boulder, 430 UCB, Boulder, CO 80309-0430, [2]Department of Applied Mathematics, University of Colorado at Boulder, 526 UCB, Boulder, CO 80309-0526, [3]Department of Molecular, Cellular and Developmental Biology, University of Colorado at Boulder, 347 UCB, Boulder, CO 80309-0430, [4]Department of Chemistry and Biochemistry, University of Colorado at Boulder, 215 UCB, Boulder, CO 80309

TABLE OF CONTENTS

1. Abstract
2. Introduction
3. Methods
4. Results
5. Discussion
6. Acknowledgments
7. References

1. ABSTRACT

The abundance of simple but functional RNA sites in random-sequence pools is critical for understanding emergence of RNA functions in nature and in the laboratory today. The complexity of a site is typically measured in terms of information, i.e. the Shannon entropy of the positions in a multiple sequence alignment. However, this calculation can be incorrect by many orders of magnitude. Here we compare several methods for estimating the abundance of RNA active-site patterns in the context of *in vitro* selection (SELEX), highlighting the strengths and weaknesses of each. We include in these methods a new approach that yields confidence bounds for the exact probability of finding specific kinds of RNA active sites. We show that all of the methods that take modularity into account provide far more accurate estimates of this probability than the informational methods, and that fast approximate methods are suitable for a wide range of RNA motifs.

2. INTRODUCTION

Our understanding of the evolution of functions in DNA, RNA and protein sequences rests critically on the probability of finding sequences with specific functions by chance in collections of random sequences. Of particular importance is the probability of calculating the abundance of RNA active sites in short sequences, as longer sequences become increasingly improbable in primitive conditions. Although we have very good models for understanding evolution of a set of related, or homologous, sequences from a common ancestor according to Markov models of evolution (1), our understanding of the probability of sequences arising independently in different groups of organisms is far more limited. In order to fully understand the evolutionary processes leading to a set of functional sequences, whether produced by natural selection over billions of years, or by artificial selection in a few weeks in the laboratory, we must develop methods for assessing whether it is more probable that a given collection of

sequences arose once (and it was passed on with modifications through successive generations), or arose many times. This problem is especially important in RNA, which is a model for molecular evolution in the laboratory and which may have preceded both DNA and protein in an 'RNA World' stage of evolution, in which RNA acted both as catalyst and genetic material (2).

The last 25 years have brought a revolution in RNA biology, with the recognition that RNA can play important catalytic and regulatory roles in the cell rather than just being a passive messenger. Of particular importance is a laboratory technique called SELEX, or *in vitro* selection, in which random-sequence RNA libraries are synthesized and screened for various properties (3-5). SELEX has produced RNAs that can perform many functions relevant to the origin of modern metabolism: there are dozens of examples, including amino acid binding (6-12), nucleotide synthesis (13), and self-aminoacylation (14, 15). Artificially selected RNAs can also bear striking resemblance to natural systems. For example, an RNA selected to bind a transition state analog of the peptidyl transfer reaction contains a conserved 8-base sequence that is identical to a conserved 8-base sequence in the ribosome at the site that naturally performs this reaction in all cells during translation (16, 17), and artificially selected RNAs that bind amino acids recapture properties of the canonical genetic code (18-20).

One striking feature of both naturally and artificially selected RNAs is that they are highly modular (21, 22). In other words, they tend to consist of short conserved pieces of the active site (the 'modules') that are separated by essentially random regions of sequence (the 'spacer'). Modular RNAs consist of specific sequence motifs in the context of specific secondary structure elements. For example, the minimal tryptophan-binding site consists of a CYA opposite a GAC in an internal loop flanked by helices (12), i.e. two modules each consisting of three conserved bases and flanked by several base pairs on either side. However, essentially any sequence can occur between the two halves of the helices. This modularity is important because both natural and artificial selection (SELEX) recover motifs of this form: modular motifs have a combinatorial advantage over single-module motifs, so should be isolated more often if they are stable. Another important feature of RNA motifs is that they are held together by base pairing, leading to correlations in the sequence (e.g. in a base-paired region, if we have a C at one location, we must have a G at the location that pairs with it).

It is important to calculate the probability of finding a given RNA in a random-sequence background for several reasons. First, we can estimate how likely a particular sequence would be observed in a SELEX experiment, and perhaps tune random-sequence pools to maximize the probability of occurrence of interesting motifs (23-25). Second, we have very good methods for estimating the probability of obtaining a set of sequences given an evolutionary model (26-29), but the probability of obtaining the sequences through multiple origins (30) is not well understood. For example, we know for certain that the hammerhead ribozyme has evolved at least three times: at least once in nature, and at least twice in artificial selections in the Breaker and Szostak lab (31, 32). Improving our estimates of the probabilities of modular RNA sites can help us understand whether different RNAs that contain the same motif most likely had a common ancestor or evolved independently. Third, genomewide searches for motifs, such as those performed by the Infernal package (33) and used in the Rfam database (34) return many matches, and we thus need to calculate the statistical significance of a given motif to rule out the null hypothesis that it evolved by chance. Fourth, understanding the regions of nucleotide composition that make RNA functions most likely may provide clues about which genomes are most likely to evolve which functions, and about the chemical conditions under which the RNA world might have emerged (35). For example, we might be able to address unsolved problems such as why some bacteria use RNA for regulation where others use proteins. Perhaps genome composition, which varies over a huge range, favors formation of riboswitches in species that have the right composition. Similar considerations might apply to the use of the hairpin, hammerhead, and HDV self-cleaving motifs, which, along with many other self-cleaving RNAs, perform similar functions (32).

Several methods have been proposed for calculating the probability of finding a correlated modular motif (referred to as 'the motif' in the text below). We note that these methods cover only the first step in calculating the probability of obtaining an active molecule: the second step is to calculate the probability of correct folding given that the sequence elements required for a motif are present, as in (25), and the final step is to calculate the probability that the molecule is functional given that it contains the sequence elements required for activity and is predicted to fold correctly, which can be achieved by laboratory experiments. However, because these probabilities are multiplied together to get the overall probability of function, errors in the first step are propagated throughout the calculation. These methods for calculating the first step, the probability of obtaining the correct sequence elements, are:

1. Information content, as used in e.g. (36, 37): in this method, a multiple sequence alignment is constructed, and examined for conserved positions, which contain the same nucleotides at corresponding sites in different sequences. The information content in bits is given by Shannon's formula (38): $I = -\Delta H$, where $H = -\sum p_i \log_2 p_i$, summed across the nucleotides in the sequence. The intuition here is that in a random RNA sequence, there are 4 possible states at each position, so if the bases are equiprobable there is a reduction of 2 bits of uncertainty if only one of the four choices is acceptable ($H_{before} = -4 \times 0.25 \times \log_2 0.25 = -4 \times 0.25 \times -2 = 2$, $H_{after} = 0$, so the difference is 2 bits). Thus we have 2 bits of information per conserved nucleotide, and 2 bits of information per conserved base pair (or 1.47 bits of information if wobble pairing is allowed, because then the final uncertainty is 6/16 rather then 4/16 to account for the G-U and U-G pairs: the standard Watson-Crick pairs are A-

U, U-A, C-G, and G-C). Although the simplicity of this method is appealing, it assumes that all of the sequences are drawn from a single starting sequence, with the conserved sites appearing at specific positions within this reference frame. Converting bits to probabilities, this model implies that each conserved nucleotide or Watson-Crick base pair multiplies the probability of occurrence by 1/4 (for Watson-Crick pairs, 4/16 = 0.25 of the possible choices of two nucleotides are valid pairs), and that each conserved wobble pair multiplies the probability of occurrence of the motif by 6/16. The appeal of the Shannon formula arises from its simplicity, and from the fact that the information content of a motif is independent of the number of modules that it is broken into and of the length of the sequence in which it is embedded. However, as we shall see, these simplifying assumptions lead to substantial inaccuracies in calculation as indeed they are independent of the number of spacers. This method also assumes that the four bases are equiprobable, which is often reasonable in SELEX because the incorporation rates can be controlled during chemical synthesis, but is not reasonable in genomes where we know the background base frequencies vary widely. This method also assumes that there are no correlations among successive positions in the sequence, i.e. that the base at each position does not affect the frequencies of the bases that follow.

2. Poisson approximation across sites (22): in this method, we calculate the probability of observing the motif in a single trial (i.e. of finding it in a single random sequence of the precise length of the motif). We then calculate the number of ways to place the modules of the motif within the longer sequence, and use the Poisson formula to calculate the probability of zero occurrences in the number of 'trials' corresponding to the sequence. The complement of the probability of the zero class is the probability that the motif occurred at least once. This method assumes that each possible match location is independent, and that a match is extremely unlikely, and essentially calculates the probability of a match anywhere in the sequence. Although the assumption of independence may lead to reasonable approximations in relatively long sequences (as compared to the original motif), modularity makes matches much more likely, thus violating one of the key assumptions justifying the Poisson approximation (39). Indeed, we shall see that highly modular motifs can lead to less accurate estimates on the probability of finding the motif when the probability of occurrence of the individuals are large. Like the information method, this method assumes that there are no correlations among successive positions in the sequence.

3. State machine/transition matrix (40): unlike the Poisson approximation, this method provides an exact way of calculating the probability of occurrence of a modular motif, although it is very expensive computationally. In this method, we embed the random sequences into a deterministic finite automaton (finite state machine) that detects matches with each possible pattern that could lead to the occurrence of the motif in the sequence (see Methods below for additional details). When we embed an i.i.d. (independent and identically distributed) random sequence of RNA bases into the automaton, the resulting stochastic process is a first-order homogeneous Markov chain on the states of the automaton. The probability that the motif is present in the random-sequence is then equivalent to the probability that the Markov chain is absorbed into a state indicative of a match with the motif. This probability can be calculated in two ways. First, the probability transition matrix of the chain, which gives the probability of moving from each state to each other state, may be exponentiated to the length of the sequence, and the entries associated with matches with the motif may be summed up to determine the probability of a match. Second, a network flow approach, in which we simulate each additional character by visiting each state that can be reached in a given number of characters, multiplying the probability of that state by the probability of each of the possible characters, and adding the result to the probability of the state that is reached by adding a character. In practice, both methods give very similar results when both can be implemented, but differ substantially in run time and computer memory usage. As we shall see, the automata-based approach becomes computationally infeasible in both memory and CPU time with very small numbers of correlations (i.e. base pairs). The most complex case is when correlations occur between modules (i.e. the modules are base paired to one another), which forces us to consider the product of the automata associated with each of the unique and simpler patterns that build up the motif (in our case, a concatenation of Aho-Corasick automata (41)). We omit results from the network flow approach in what follows because the method is more heuristic than mathematical, and does not provide substantially different results from the inclusion-exclusion approach below (on which we can place more precise bounds).

4. Inclusion-exclusion approach: this method can be used both for exact calculations of the probability of occurrence of a modular motif (for small cases), or to give bounds for this probability (for larger cases). In this method, we also use product automata as above. However, we aim for bounds of the same order of magnitude rather than an exact calculation of the probability of observing the motif. Using the inclusion-exclusion formula (42), we can calculate the probability of occurrence of the motif in a random-sequence by determining the probability of match-ing individual, pairwise, three-way, etc. combinations of the unique patterns that build up the motif. Combining this probability as P(individual)−P(pairwise)+P(three-way)-... we can recover the exact probability of matching the motif. However, according to Bonferroni's inequality (42), this exact probability is bounded by any two successive partial sums occurring in the inclusion-exclusion formula. In particular, if for a small k, the first k terms in the inclusion-exclusion formula provide a tight estimate of the probability, the associated bounds in the Bonferroni's inequality will be of the same order, and we only need to consider the product of at most *k* automata to obtain a tight approximation for the probability of matching the motif. When there are too many combinations of the *k*

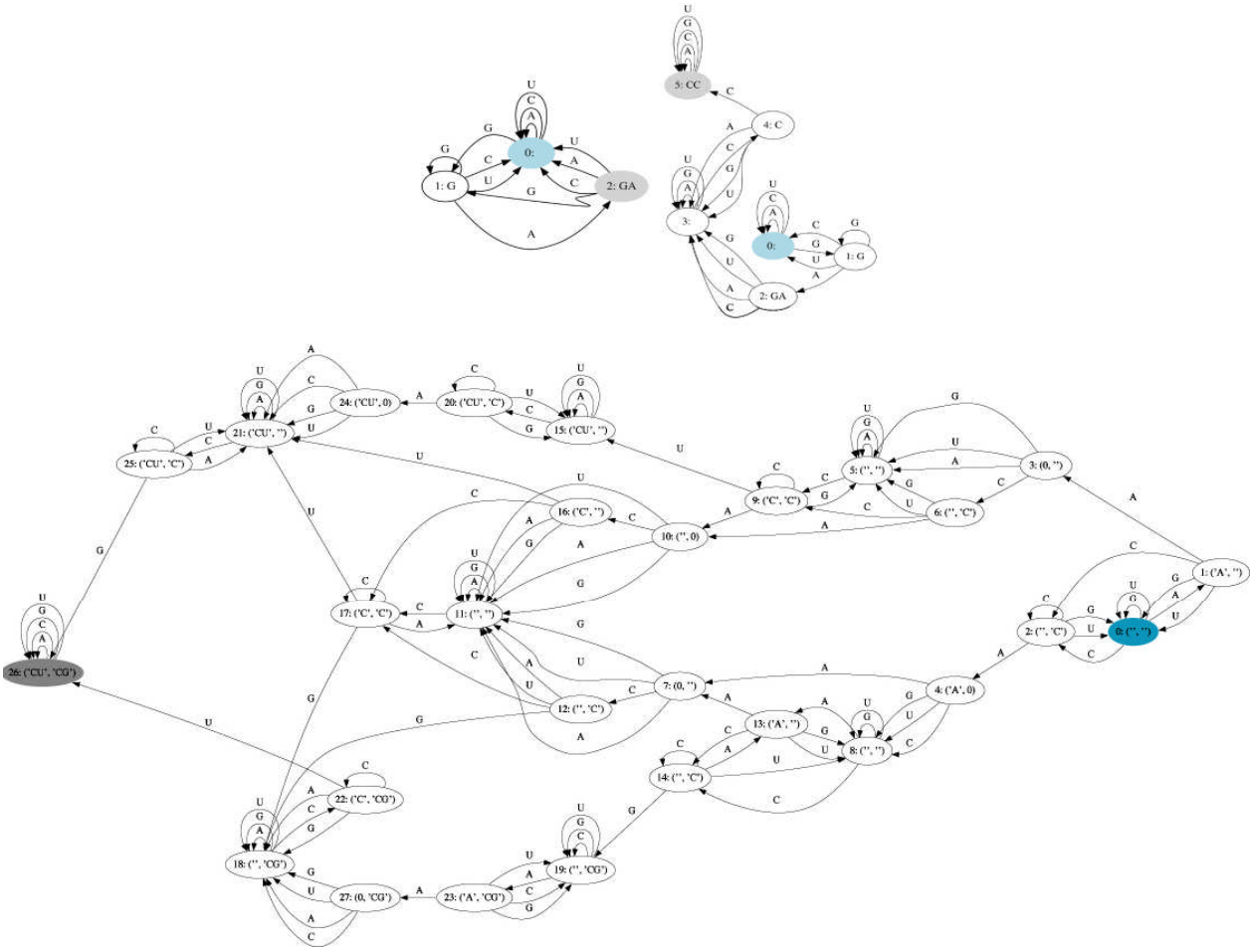

Figure 1. Auxiliary deterministic finite automata needed for detecting at least one match with the correlated modular motif $(NA*_1CN)$ in a sequence of RNA nucleotides. Top-left, Aho-Corasick automaton that recognizes any match with the keyword GA in an RNA sequence. State 0: is the initial state. Visits to state 2:GA correspond to matches with the keyword GA within the RNA sequence. Top-right, automaton that recognizes the motif $GA*_1CC$ i.e. the motif $(NA*_1CN)$ when the correlation is replaced with the pair GC. State 0: is again the initial state. Up to minor modifications states 0:, 1:G and 2:A correspond the the automaton on the left, whereas states 3:, 4:C and 5:CC correspond to the states of the Aho-Corasick automaton that detects matches with the keyword CC. State 2:A is visited for the first time when GA is first encountered in the RNA sequence. Transitions from state 2:A to state 3: represent the unconstrained region of at least one nucleotide in the motif $GA*_1CC$. State 5:CC may only be visited from state 4:C once the keyword CC is detected. Absorption into this state guarantees a match with the uncorrelated motif $GA*_1CC$. Bottom, product automaton for detecting at least one match with the correlated modular motif $(NA*_1CN)$ when the correlation N is restricted to the value A or C. States are ordered pairs of the form (v_1,v_2), with v_1 a state in the automaton that detects at least one match with the motif $AA*_1CU$, and v_2 a state in the automaton that detects at least one match with the motif $CA*_1CG$. States labeled with the prefixes 15:, 20:, 21:, 24: and 25: guarantee a match with $AA*_1CU$ without a match with $CA*_1CG$. States labeled with the prefixes 18:, 19:, 22:, 23: and 27: guarantee a match with $CA*_1CG$ but without a match with $AA*_1CU$. The state labeled with the prefix 26: guarantees a match with both $AA*_1CU$ and $CA*_1CG$. The automaton that detects at least one match with the motif $(NA*_1CN)$ would require the product of four automata and it is not displayed in here due to limitations of space. All the plots were obtained using the software Graphviz, available at graphviz.org.

unique patterns that build up the motif, we can obtain asymptotic confidence intervals for the terms appearing in Bonferroni's inequality via Monte Carlo simulation. As we shall see, this new method provides accurate estimates even for highly modular motifs for which the use of the Poisson approximation is badly justified and for which the state machine/transition matrix approach is completely impractical. One important feature of this approach is that it can be extended to sequences with memory (i.e. correlations among successive positions). These correlat-

ions have been observed in genomes, and are widely used for gene-finding (43).

In this paper, we test the accuracy of these different methods on motifs of different length and composition in different random-sequence backgrounds.

3. METHODS

We used an exact and a stochastic version of the inclusion-exclusion method to either estimate, or to find a $100(1-\alpha)\%$ asymptotic confidence interval for the probability p of observing a given motif in a random sequence of l RNA bases (nucleotides) produced in a SELEX experiment. We typically consider $\alpha=0.01$, or $\alpha=0.05$.

The range of motifs considered here cover many of the small motifs routinely found in SELEX. The motifs are correlated (i.e. they contain base pairs), are composed of either one, two or three modules, and have constant regions of various lengths. For example, the motif (N(NACGUACGUAC$*_1$GU$*_1$AN)N) consists of three modules, two correlations (base pairs), and a constant region totaling thirteen nucleotides. Modules are separated by unconstrained regions. The notation $*_1$ refers to an unconstrained region of at least one nucleotide. The modules of this motif are (N(NACGUACGUAC, GU and AN)N), where the N's represent bases that may be any of the four nucleotides. The two bases directly within a pair of parentheses are correlated and must pair with each other.

When the motif consists of n correlations, the probability p corresponds to the probability that either of $m = 4^n$ simpler motifs (i.e. uncorrelated motifs) is present in the random sequence. In the example discussed above, n=2, m=16 and one of the uncorrelated motifs is CAACGUACGUAC$*_1$GU$*_1$AUG, in which the outer correlation was replaced by the pair of bases CG and the inner by AU.

Each of the uncorrelated motifs may be identified in a non-random sequence of nucleotides using a deterministic finite automaton (44). We construct such an automaton by concatenating the Aho-Corasick automata (41) associated with each of the constant regions. Except for the automaton associated with the last constant region, all transitions from the terminal state of an Aho-Corasick automaton are redefined so as to lead to the initial state of the next. The terminal state of the Aho-Corasick automaton associated with the last constant region is, however, reset to be an absorbent state (an absorbent state is a state that always returns to itself when additional characters are fed into the automaton). The resulting automaton will have as many as #(nucleotides in the constant region)+#(modules) number of states. (See Figure 1 for a more detailed explanation of these constructions for a motif consisting of two modules, one correlation and a constant region of two nucleotides.)

When m is of a manageable size we may consider the product of the automata associated with each of the uncorrelated motifs to determine p directly. The states of this automaton are ordered m-tuples of the form $(v_1,\ldots,v_m)$, where v_i is always a state in the automaton associated with the i-th uncorrelated motif. (In principle this automaton may have at most {#(nucleotides in the constant region)+#(modules) }^{m} states. In many situations of interest, this upper-bound is exaggerated because not all states of the form $(v_1,\ldots,v_m)$ may be reached from the initial state of the product automaton i.e. the state $(q_1,\ldots,q_m)$, with q_i the initial state of the automaton associated with the i-th uncorrelated motif.) By embedding a random sequence of i.i.d. nucleotides into this product automaton we are guaranteed to obtain a first-order homogeneous Markov chain (40). Indeed if p_β denotes the proportion of base $\beta \in \{A,C,G,U\}$ used in the SELEX experiment then the probability transition from a state s_1 into a state s_2 is $\sum_\beta p_\beta$, where an index β is accounted for in this summation if and only if there is a transition from state s_1 into state s_2 labeled with the character β. If P denotes the probability transition matrix of the resulting Markov chain and $P^l(s_1,s_2)$ denotes the entry associated with row s_1 and column s_2 of the power matrix P^l then

$$p = \sum_s P^l(q,s)\ ,$$

where q is the initial state of the product automaton, and the indices s are restricted to be those states of the form $(v_1,\ldots,v_m)$ where at least one of the entries v_i is a terminal state.

Unfortunately, in most situations of interest, the product automata described above do not scale well. It is for these cases that an estimate of p rather than an exact formula may be more suitable. If we denote by E_i the event that the i-th of the uncorrelated motifs is present in a random sequence of length l, it follows from the inclusion-exclusion formula (42) that

$$p = S_1 + S_2 + \ldots + S_m, \qquad (1)$$

where

$$S_1 = \sum_i \mathrm{Prob}(E_i);\quad S_2 = -\sum_{i<j} \mathrm{Prob}(E_i \cap E_j);$$

$$S_3 = \sum_{i<j<k} \mathrm{Prob}(E_i \cap E_j \cap E_k);\quad \textit{etc.}$$

In general, if $|I|$ denotes the number of indices in the set I we have that

$$S_k = (-1)^{k+1} \cdot \sum_{I\subset\{1,\ldots,m\};|I|=k} \mathrm{Prob}\left(\bigcap_{i\in I} E_i\right).$$

To obtain S_m we need to compute the probability that all the uncorrelated motifs are present in the random sequence of nucleotides. To the best of our knowledge this

can only be attempted by means of the product automaton described in the above paragraph. In particular, when this automaton does not scale properly, the use of formula (1) is also impractical.

Observe that the summation S_1 consists of m terms, S_2 of $m(m-1)/2$ terms, and S_3 of $m(m-1)(m-2)/6$ terms. In general S_k will consist of about $m^k/k!$ terms when m is large. Furthermore, if the first $2D$ sums in (1) may be computed exactly, we may use Bonferroni's inequality (42) to obtain that

$$\sum_{k=1}^{2D} S_k \le p \le \sum_{k=1}^{2D-1} S_k \ .$$

We will refer to the above as the Bonferroni's inequality of depth D. The above inequality is useful when the lower- and the upper-bound are of the same order. Unfortunately, it is not in general true that the larger D, the tighter are the two bounds above. Because of this, the inequality

$$\max_{d=1,..,D} \sum_{k=1}^{2d} S_k \le p \le \min_{d=1,..,D} \sum_{k=1}^{2d-1} S_k \qquad (2)$$

provides the best bounds for p when all the sums involved in the Bonferroni's inequality up to depth D may be computed exactly.

To compute the sum S_k we need, for each possible combination of k uncorrelated motifs of the m motifs, to compute the probability that all the k uncorrelated motifs are present in the random sequence of l nucleotides. This can be performed similarly to the method discussed earlier, but considering only the product of the automata associated with the k uncorrelated motifs. The resulting automaton will have a single absorbing state. Furthermore, the embedding of the random sequence into this automaton is again guaranteed to be a first-order homogeneous Markov chain (40). The probability that all the k uncorrelated motifs are present in a random sequence of length l is equal to the probability that the Markov chain is absorbed within the first l steps, when started at the initial state. This probability can be computed as $P^l(q,t)$, where q is the initial state of the product automaton and t its unique absorbent state.

For relatively small values of k each of the probabilities appearing in S_k may be computed exactly using the method described above. This is because the automata associated with the uncorrelated motifs scales linearly with the total length of the constrained regions and the number of modules (which are quantities independent of the number of correlations!). The only issue here is that when m is large it may take an infeasible amount of time to determine exactly all of the probabilities appearing in S_k. In this situation it is advisable to use Monte Carlo simulation to find an estimate $\hat{S}_k$ of S_k for the first few values of k. Because, after the proper re-scaling, $(\hat{S}_k - S_k)$ is approximately a standard Gaussian distribution, we may obtain asymptotic confidence intervals and/or asymptotic upper- or lower-confidence bounds for S_k. For example, by considering the Bonferroni inequalities up to depth one and two it follows from (2) that

$$(S_1 + S_2) \le p \le \min\{S_1, S_1 + S_2 + S_3\} \quad (3)$$

We can find a $100(1-\alpha)\%$ asymptotic confidence interval for p using the following procedure:

(a) Independently determine a $\sqrt[4]{1-\alpha}$ asymptotic lower-confidence bound L_1 for S_1, and also a $\sqrt[4]{1-\alpha}$ asymptotic lower-confidence bound L_2 for S_2. Then $(L_1 + L_2)$ is at least a $\sqrt[2]{1-\alpha}$ asymptotic lower-confidence bound for $(S_1 + S_2)$, i.e., the event $(L_1 + L_2) \le (S_1 + S_2)$ has approximately a probability of at least $\sqrt[2]{1-\alpha}$.

(b) Independently, and following the same logic as above, determine a $\sqrt[4]{1-\alpha}$ asymptotic upper-confidence bound U_1 for S_1, and also a $\sqrt[4]{1-\alpha}$ asymptotic upper-confidence bound U_3 for $(S_1 + S_2 + S_3)$. Then $\min\{U_1, U_3\}$ is a $\sqrt[2]{1-\alpha}$ asymptotic upper-confidence bound for the $\min\{S_1, S_1 + S_2 + S_3\}$, i.e., the event $\min\{S_1, S_1 + S_2 + S_3\} \le \min\{U_1, U_3\}$ has approximately a probability of at least $\sqrt[2]{1-\alpha}$.

(c) Due to (3), it follows from (a) and (b) that the event $(L_1 + L_2) \le p \le \min\{U_1, U_3\}$ will occur with a probability of approximately $1-\alpha$.

A similar approach can be implemented using the more robust inequality

$$\max\{S_1 + S_2, S_1 + S_2 + S_3 + S_4\} \le p \le \min\{S_1, S_1 + S_2 + S_3\} \quad (4)$$

One key consideration for using Monte Carlo simulation to estimate the bounds in (3) or (4) is that only a few of the probabilities appearing in S_k may be needed to obtain good estimates. For relatively small values of k this is time efficient because the computation of each of the probabilities appearing in S_k requires low time and low memory complexity. Instead if we were to consider all of the about $m^k/k!$ probabilities appearing in S_k then we would obtain an exact estimate for S_k at the expense of extensive computation.

All methods described here, as well as the previously published methods outlined in the introduction, were implemented in Python.

4. RESULTS

In almost all cases, the automata-based methods agreed very well with one another and with the Poisson approximation, and disagreed greatly from the information-based method. These discrepancies increased as the mod-ularity of the sequences increased. Figure 2 shows results for the single-module case on the 2-letter alphabet. All three methods that take modularity into account agree almost exactly (the lines are superimposed), but the information-based method (blue line) is different by about an order of magnitude. As expected, the information method does not vary with the length of the spacer, whereas all other methods produce probabilities that increase rapidly as the amount of spacer increases. The run-time performances of the Poisson and information content methods were comparable and were much faster than the automata-based methods, ranging from 2 to 9 orders of magnitude faster in the range covered by the experiments shown here. (The computation time for the Poisson and information methods is essentially constant at less than a thousandth of a second for this range of sizes, whereas the computation time for the automata-based methods rapidly increased to several tenths of a second (for variation in the constant region or spacer, with two base pairs) or several minutes (for increasing the correlations) as the problem sizes increased. Note that the approximate inclusion-exclusion approach levels off because only a fixed number of samples is evaluated each time, no matter how many possible combinations there are.) Confidence intervals are not shown on these plots because they are so tight as to be undetectable: the average error was much less than 1% of the estimated value, and the maximum error we observed in these calculations was 3.5%. In cases where there were fewer than 16 possibilities (4 correlations for the 2-character alphabet, 2 correlations for the 4-character alphabet), the confidence intervals were less than the machine precision.

When the results are extended to more modular sequences, the information content method continues to differ greatly from all other methods, and the discrepancy increases as the number of modules increases from 1 to 3. Figure 3 shows probability estimates for 2 and 3 modules (parameter settings otherwise the same as in Figure 2). Run-time performance was comparable to the 1-module case (data not shown). The Poisson approximation tends to overestimate the probability by about a factor of 2 in cases where the probability is high or the number of correlations is large, consistent with the violation of the independence of trials assumption and rarity of matches that the approximation makes (recall that the transition matrix method produces probabilities that are exact to within machine precision).

For the 4-character alphabet (simulating RNA directly), even two correlations increase the dimensionality of the problem so much that it is impossible to calculate the probabilities using the exact method on machines with 2−8 GB of memory. Figure 4 shows results for the other methods considering a single module. In these cases, the Poisson and approximate inclusion-exclusion method agree very well with one another. In all cases where the probability of each module within the motif is below 0.01 in a single trial, the Poisson gives essentially quantitative agreement with the inclusion-exclusion approach ($r^2 > 0.9999$ over 13 orders of magnitude, for both equal nucleotide frequencies as shown in the plot and for unequal nucleotide frequencies).

5. DISCUSSION

The automata-based approaches in principle provide an exact method of calculating the probability of observing a given modular pattern in a longer sequence, and hence the probability of observing a given RNA motif in random-sequence pools. Because the exact calculations scale poorly from a 2-letter alphabet to the 4-letter RNA alphabet, however, we must use approximate techniques. The approach to calculating confidence intervals that we describe here, which to our knowledge has not previously been applied to nucleic acid patterns, allows us to calculate for the first time the precise amount of error we expect these approximate calculations to introduce. The results demonstrate that in most cases, except when the probability of individual modules is large, the Poisson approximation provides almost quantitative agreement with the true probability. In contrast, the information approach provides estimates that are many orders of magnitude different from the true probabilities.

These results demonstrate that the Poisson approximation should always be used in cases where the individual modules are rare, as it is at least as fast to calculate as the information-based method and provides much more accurate results. However, in cases where there are many modules, and where some of those modules are extremely short or degenerate, the approximate automata-based approach described here can be used to calculate the probabilities almost exactly and to place confidence intervals that set bounds on the region within which the probabilities must lie as described above. These calculations depend solely on a precise definition of the degeneracy of the motif such as could be obtained with a few dozen to a few hundred unique sequences (a number typically collected in laboratory experiments), although practical considerations such as amplification bias may limit the agreement of these calculations with experimental results.

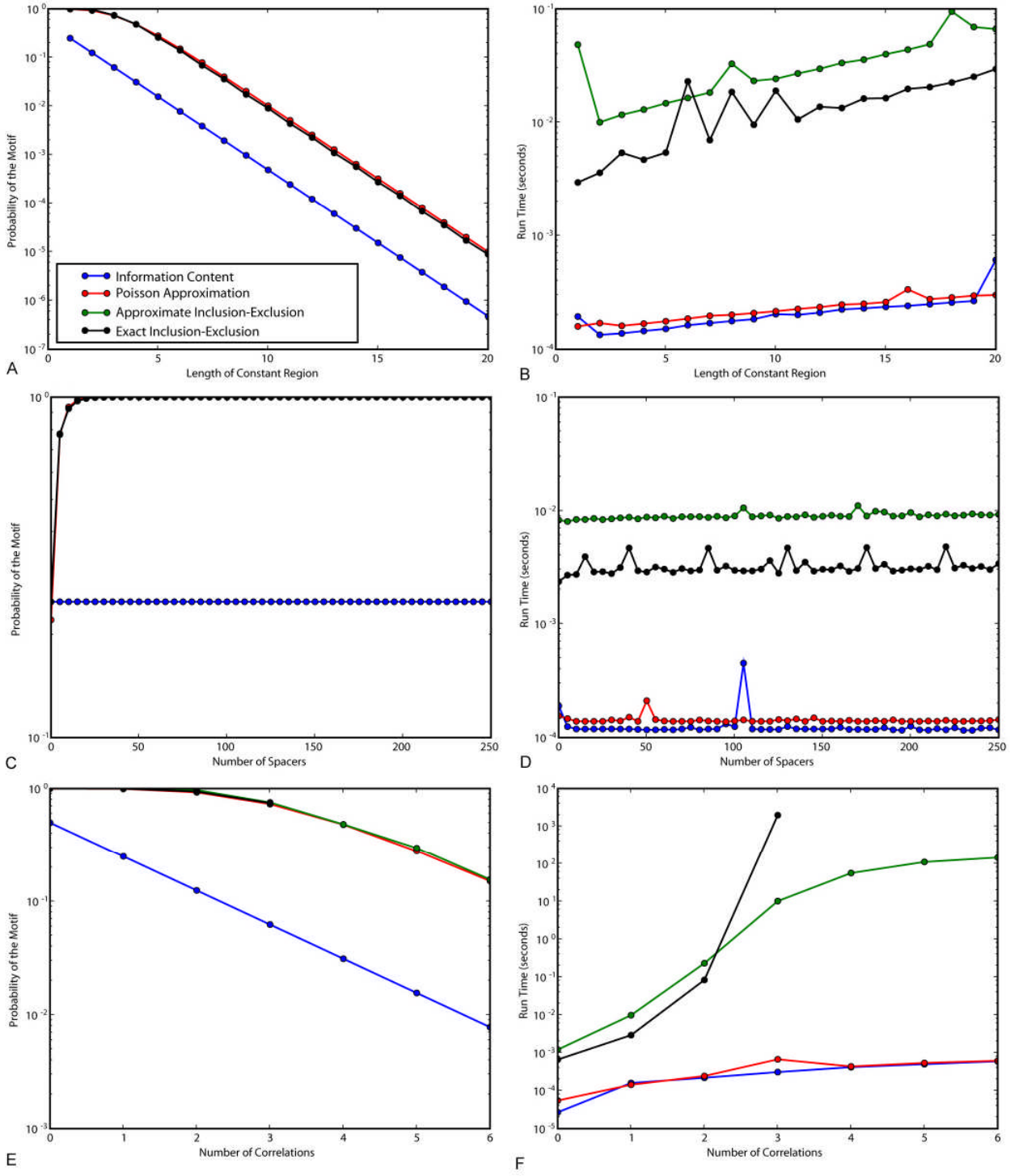

Figure 2. Effects of varying the length of the constant region, amount of spacer, and number of correlations using the 2-letter alphabet in the case of a single module. (a) Probability of motif against length of constant region. (b) Run time against length of constant region. (c) Probability of motif against amount of spacer. (d) Run time against amount of spacer. (e) Probability against number of correlations. (f) Run time against number of correlations (note that exact inclusion-exclusion could not be calculated for >3 correlations due to time constraints). Standard settings were 1 correlation, 1 base per module, and 20 bases of spacer.

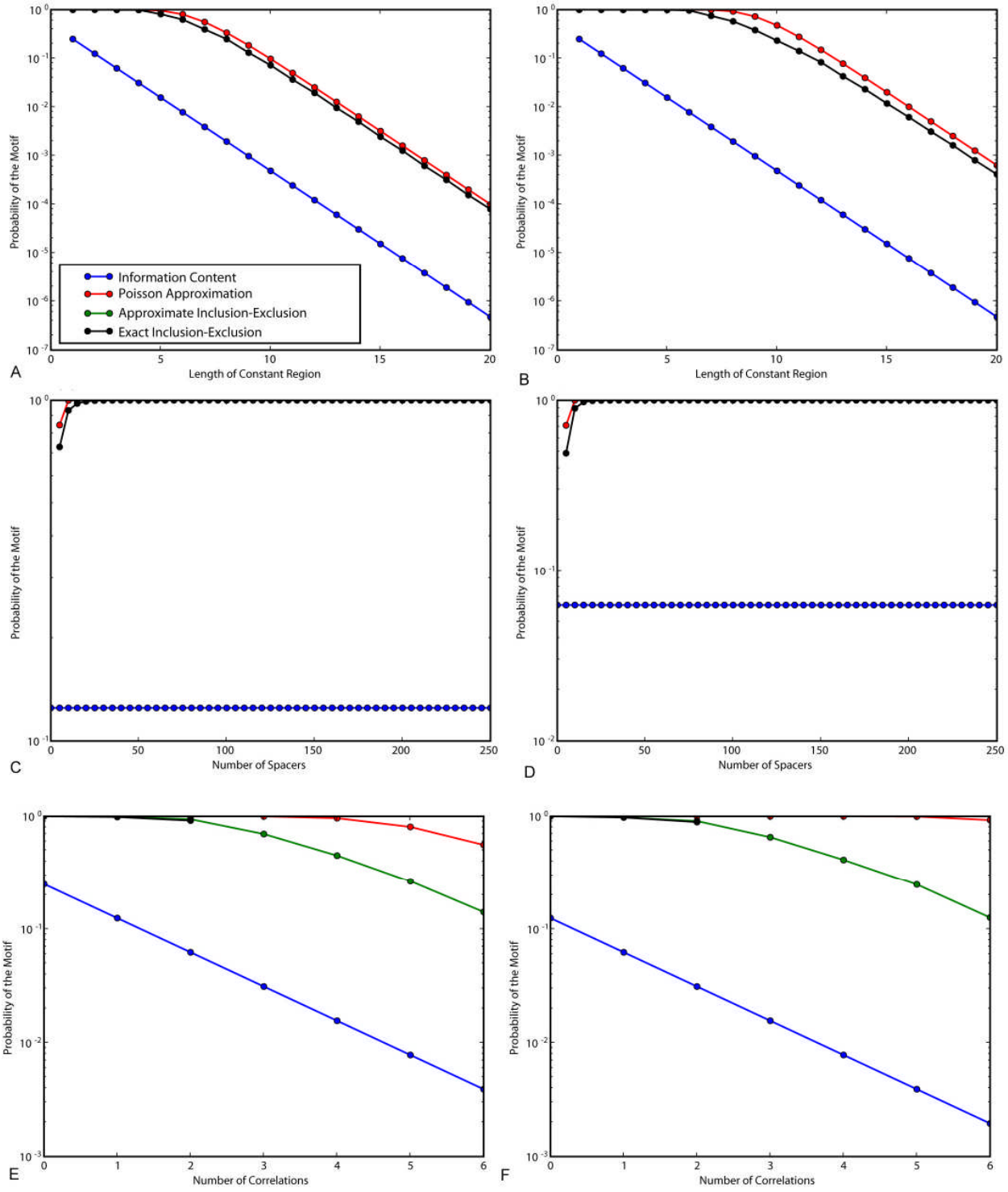

Figure 3. Effects of varying the length of the constant region, amount of spacer, and number of correlations using the 2-letter alphabet on the probability of finding a two-module or a three-module motif. (a) Two modules, varying length of constant region. (b) Three modules, varying length of constant region. (c) Two modules, varying length of spacers. (d) Three modules, varying length of spacers. (e) Two modules, varying number of correlations. (f) Three modules, varying number of correlations. Standard settings were 1 correlation, 1 base per module, and 20 bases of spacer. Note that the black and green lines for the two inclusion-exclusion methods are in identical locations on most plots, and are superimposed.

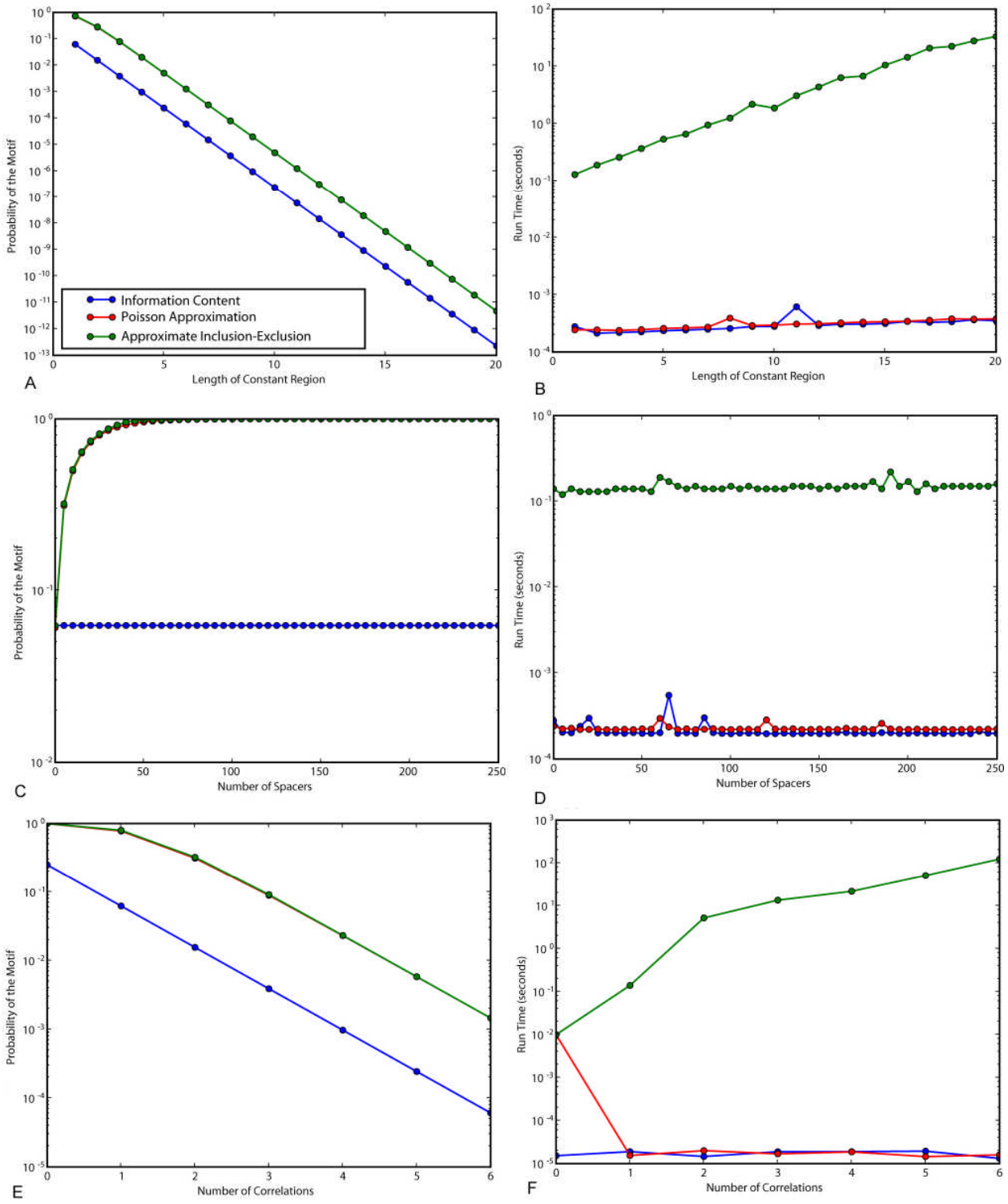

Figure 4. Effects of varying the length of the constant region, amount of spacer, and number of correlations using the 4-letter alphabet in the case of a single module. (a) Probability of motif against length of constant region. (b) Run time against length of constant region. (c) Probability of motif against amount of spacer. (d) Run time against amount of spacer. (e) Probability against number of correlations. (f) Run time against number of correlations. Standard settings were 1 correlation, 1 base per module, and 20 bases of spacer.

These calculations have direct pragmatic implications for studies of RNA active sites. In previous work (21, 22, 25), we demonstrated that the modularity of motifs was extremely important for determining their frequency of occurrence (using the Poisson and related approaches). In later work, we demonstrated that the different modularity of two RNA active sites for binding isoleucine switched orders depending on the length of the random region, in accordance with the Poisson calculations (45). The present work demonstrates that these conclusions based on the Poisson approximation are almost certain to be within a few percent of the true probabilities, and that the methodology can be applied with confidence to compute the probabilities of a wide range of RNA active sites in different random-sequence pools.

The information approach to determining the complexity of RNA active sites should be discouraged, as the alignment assumption introduces a subtle but profound bias that results in estimates of the probability that are incorrect by orders of magnitude. Worse, the failure of the information-based approach to account for the number of modules into which the motif is divided can result in inversion of the relative probabilities of two motifs, and therefore incorrectly lead us to suspect that selection for or against one of the motifs is at work when the observed abundances are consistent with the true probabilities but not with the information calculations. Even in cases in which the Poisson approximation was inaccurate, we found that the only case where the information method outperformed the Poisson approximation was when the probability of finding the motif was very high and there was only one position for it in the longer sequence.

As our previous work (21, 22) showed, simple RNA motifs are extremely abundant, especially in longer sequences, although the probability of occurrence decreases rapidly as the length of the constant region and the number of base pairs increases. Including the paired regions is essential for estimating the abundance accurately, as per the cautionary note in (21), so the calculations given in (25) are to be preferred.

The implications of these results for understanding RNA motifs are profound. RNA motifs cannot be assigned intrinsic complexity because the probability of occurrence of the motif depends on the length and composition of the sequence in which it is embedded. Thus, a more modular motif may have a lower probability of occurrence than a less modular motif in a SELEX experiment performed with a short random region, but a higher probability of occurrence when the same SELEX experiment is repeated with a longer random region. However, the ability to calculate motif probabilities accurately greatly improves our ability to interpret the results of these experiments, and may pave the way for understanding how many times particular kinds of RNA sites evolved in nature, and what RNA abundance complex motifs would first appear.

6. ACKNOWLEDGMENTS

We thank Micah Hamady for running the large-scale computations on the Keck RNA Bioinformatics Facility at CU Boulder, and for setting up and maintaining this facility, and Catherine Lozupone for helpful comments on a draft of the manuscript. This work was supported in part by the National Institutes of Health Supplemental Biocomplexity award 3R01GM048080-13S1. M.E.Ll. and R.K. contributed equally to this work.

7. REFERENCES

1. Kimura, M.: Neutral Theory of Molecular Evolution. *Cambridge University Press*, Cambridge (1983)
2. Gilbert, W.: The RNA World. *Nature*, 319, (1986)
3. Ellington, A. D. & J. W. Szostak: *In vitro* selection of RNA molecules that bind specific ligands. *Nature*, 346, 818-22 (1990)
4. Robertson, D. L. & G. F. Joyce: Selection *in vitro* of an RNA enzyme that specifically cleaves single-stranded DNA. *Nature*, 344, 467-8 (1990)
5. Tuerk, C. & L. Gold: Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science*, 249, 505-10 (1990)
6. Connell, G. J. & M. Yarus: RNAs with dual specificity and dual RNAs with similar specificity. *Science*, 264, 1137-41 (1994)
7. Geiger, A., P. Burgstaller, H. von der Eltz, A. Roeder & M. Famulok: RNA aptamers that bind L-arginine with sub-micromolar dissociation constants and high enantioselectivity. *Nucleic Acids Res*, 24, 1029-36 (1996)
8. Illangasekare, M. & M. Yarus: Phenylalanine-binding RNAs and genetic code evolution. *J Mol Evol*, 54, 298-311 (2002)
9. Lozupone, C., S. Changayil, I. Majerfeld & M. Yarus: Selection of the simplest RNA that binds isoleucine. *RNA*, 9, 1315-22 (2003)
10. Majerfeld, I., D. Puthenvedu & M. Yarus: RNA affinity for molecular L-histidine; genetic code origins. *J Mol Evol*, 61, 226-35 (2005)
11. Majerfeld, I. & M. Yarus: Isoleucine:RNA sites with associated coding sequences. *RNA*, 4, 471-8 (1998)
12. Majerfeld, I. & M. Yarus: A diminutive and specific RNA binding site for L-tryptophan. *Nucleic Acids Res*, 33, 5482-93 (2005)
13. Unrau, P. J. & D. P. Bartel: RNA-catalysed nucleotide synthesis. *Nature*, 395, 260-3 (1998)
14. Illangasekare, M., G. Sanchez, T. Nickles & M. Yarus: Aminoacyl-RNA synthesis catalyzed by an RNA. *Science*, 267, 643-7 (1995)
15. Illangasekare, M. & M. Yarus: Specific, rapid synthesis of Phe-RNA by RNA. *Proc Natl Acad Sci U S A*, 96, 5470-5 (1999)
16. Welch, M., I. Majerfeld & M. Yarus: 23S rRNA similarity from selection for peptidyl transferase mimicry. *Biochemistry*, 36, 6614-23 (1997)
17. Yarus, M. & M. Welch: Peptidyl transferase: ancient and exiguous. *Chem Biol*, 7, R187-90 (2000)
18. Knight, R. D. & L. F. Landweber: Rhyme or reason: RNA-arginine interactions and the genetic code. *Chem Biol*, 5, R215-20 (1998)

19. Yarus, M.: RNA-ligand chemistry: a testable source for the genetic code. *RNA*, 6, 475-84 (2000)
20. Yarus, M., J. G. Caporaso & R. Knight: Origins of the genetic code: the escaped triplet theory. *Annu Rev Biochem*, 74, 179-98 (2005)
21. Knight, R. & M. Yarus: Finding specific RNA motifs: function in a zeptomole world? *RNA*, 9, 218-30 (2003)
22. Yarus, M. & R. Knight: The Scope of Selection. In: The Genetic Code and the Origin of Life. Ed: L. Ribas de Pouplana. *Landes Bioscience*, Georgetown, TX (2004)
23. Gevertz, J., H. H. Gan & T. Schlick: *In vitro* RNA random pools are not structurally diverse: a computational analysis. *RNA*, 11, 853-63 (2005)
24. Kim, N., H. H. Gan & T. Schlick: A computational proposal for designing structured RNA pools for *in vitro* selection of RNAs. *RNA*, 13, 478-92 (2007)
25. Knight, R., H. De Sterck, R. Markel, S. Smit, A. Oshmyansky & M. Yarus: Abundance of correctly folded RNA motifs in sequence space, calculated on computational grids. *Nucleic Acids Res*, 33, 5924-35 (2005)
26. Felsenstein, J.: Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol*, 17, 368-76 (1981)
27. Jayaswal, V., J. Robinson & L. Jermiin: Estimation of phylogeny and invariant sites under the general Markov model of nucleotide sequence evolution. *Syst Biol*, 56, 155-62 (2007)
28. Yang, Z.: Phylogenetic analysis using parsimony and likelihood methods. *J Mol Evol*, 42, 294-307 (1996)
29. Yap, V. B. & T. P. Speed: Modeling DNA base substitution in large genomic regions from two organisms. *J Mol Evol*, 58, 12-8 (2004)
30. Bourdeau, V., G. Ferbeyre, M. Pageau, B. Paquin & R. Cedergren: The distribution of RNA motifs in natural sequences. *Nucleic Acids Res*, 27, 4457-67 (1999)
31. Salehi-Ashtiani, K. & J. W. Szostak: *In vitro* evolution suggests multiple origins for the hammerhead ribozyme. *Nature*, 414, 82-4 (2001)
32. Tang, J. & R. R. Breaker: Structural diversity of self-cleaving ribozymes. *Proc Natl Acad Sci U S A*, 97, 5784-9 (2000)
33. Nawrocki, E. P. & S. R. Eddy: Query-dependent banding (QDB) for faster RNA similarity searches. *PLoS Comput Biol*, 3, e56 (2007)
34. Griffiths-Jones, S., A. Bateman, M. Marshall, A. Khanna & S. R. Eddy: Rfam: an RNA family database. *Nucleic Acids Res*, 31, 439-41 (2003)
35. Gutfraind, A. & A. Kempf: Error-reducing Structure of the Genetic Code Indicates Code Origin in Non-thermophile Organisms. *Orig Life Evol Biosph*, 38, 75-85 (2008)
36. Carothers, J. M., S. C. Oestreich, J. H. Davis & J. W. Szostak: Informational complexity and functional activity of RNA structures. *J Am Chem Soc*, 126, 5130-7 (2004)
37. Carothers, J. M., S. C. Oestreich & J. W. Szostak: Aptamers selected for higher-affinity binding are not more specific for the target ligand. *J Am Chem Soc*, 128, 7929-37 (2006)
38. Weaver, W. & C. Shannon: The Mathematical Theory of Communication. *University of Illinois Press*, Urbana, IL (1949)
39. Lothaire, M.: Applied Combinatorics on Words. *Cambridge University Press*, Cambridge, UK (2005)
40. Lladser, M. E., M. D. Betterton & R. Knight: Multiple pattern matching: a Markov chain approach. *J Math Biol*, 56, 51-92 (2008)
41. Aho, A. V. & M. J. Corasick: Efficient String Matching: An Aid to Bibliographic Search. *Commun A C M*, 18, 333-340 (1975)
42. Durret, R.: Probability: Theory and Examples. *Duxbury Press*, New York (2004)
43. Borodovsky, M., K. E. Rudd & E. V. Koonin: Intrinsic and extrinsic approaches for detecting genes in a bacterial genome. *Nucleic Acids Res*, 22, 4756-67 (1994)
44. Hopcroft, J. E. & J. D. Ullman: Introduction to Automata Theory, Languages and Computation. *Addison-Wesley*, Boston, MA (1979)
45. Legiewicz, M., C. Lozupone, R. Knight & M. Yarus: Size, constant sequences, and optimal selection. *RNA*, 11, 1701-9 (2005)

Key Words: automata, Markov chains, pattern matching, random strings, RNA secondary structure, Aho-Corasick automata

Send correspondence to: Rob Knight, Department of Chemistry and Biochemistry, University of Colorado at Boulder, 215 UCB, Boulder, CO 80309, Tel: 303-492-1984 Fax: 303-492-7744 E-mail: rob.knight@colorado.edu

http://www.bioscience.org/current/vol13.htm