Published online February 27, 2004

1392–1403 *Nucleic Acids Research, 2004, Vol. 32, No. 4*
DOI: 10.1093/nar/gkh291

Paradigms for computational nucleic acid design

Robert M. Dirks, Milo Lin[1], Erik Winfree[2,3] and Niles A. Pierce[3,4,*]

Chemistry Department, [1]Physics Department, [2]Computer Science and Computation and Neural Systems Departments, [3]Bioengineering Department and [4]Applied and Computational Mathematics Department, California Institute of Technology, Pasadena, CA 91125, USA

Received December 14, 2003; Revised and Accepted January 28, 2004

ABSTRACT

The design of DNA and RNA sequences is critical for many endeavors, from DNA nanotechnology, to PCR-based applications, to DNA hybridization arrays. Results in the literature rely on a wide variety of design criteria adapted to the particular requirements of each application. Using an extensively studied thermodynamic model, we perform a detailed study of several criteria for designing sequences intended to adopt a target secondary structure. We conclude that superior design methods should explicitly implement both a positive design paradigm (optimize affinity for the target structure) and a negative design paradigm (optimize specificity for the target structure). The commonly used approaches of sequence symmetry minimization and minimum free-energy satisfaction primarily implement negative design and can be strengthened by introducing a positive design component. Surprisingly, our findings hold for a wide range of secondary structures and are robust to modest perturbation of the thermodynamic parameters used for evaluating sequence quality, suggesting the feasibility and ongoing utility of a unified approach to nucleic acid design as parameter sets are refined further. Finally, we observe that designing for thermodynamic stability does not determine folding kinetics, emphasizing the opportunity for extending design criteria to target kinetic features of the energy landscape.

INTRODUCTION

Understanding how to design molecular structures is an essential step in allowing technology to interface with biology and in developing systems with increasing functional density. Nucleic acids hold great promise as a design medium for the construction of nanoscale devices with novel mechanical or chemical function (1,2). Efforts are currently underway in many laboratories to use DNA and RNA molecules for applications in patterning (3), assembly (4–6), transport, switching (7–9), circuitry (10), DNA computing (11) and DNA chips (12,13). Computational sequence selection algorithms (1,14–21) are likely to play an increasing role in exploring this new design space.

A fundamental design problem consists of selecting the sequence of a nucleic acid strand that will adopt a target secondary structure. As depicted in Figure 1a, this is the inverse of the more famous folding problem of determining the structure (and folding mechanism) for a given sequence. To attempt the rational design of novel nucleic acid structures, we require both an approximate empirical physical model and a search algorithm for selecting promising sequences based on this model. Experimental feedback on the quality of the design and the performance of the design algorithm can then be obtained by folding the molecule *in vitro*. Alternatively, if this feedback loop can be closed computationally by folding the molecule *in silico*, the quality of sequence designs could be rapidly assessed and improved before attempting laboratory validation.

In designing nucleic acid sequences, we consider the two principal paradigms illustrated in Figure 1b. Positive design methods attempt to select for a desired outcome by optimizing sequence affinity for the target structure. Negative design methods attempt to select against unwanted outcomes by optimizing sequence specificity for the target structure. A successful design must exhibit both high affinity and high specificity (14), so useful design algorithms must satisfy the objectives of both paradigms, even if they explicitly implement only one.

For some applications, it may be desirable to supplement these thermodynamic design considerations with additional kinetic requirements. For example, in designing molecular machines (8), selecting sequences that fold or assemble quickly may be crucial, since naturally occurring RNA sequences have been observed to have persistent metastable states (22) and theoretical models suggest that random sequences have highly frustrated energy landscapes with folding times that grow exponentially with sequence length (23). Alternatively, it may be important to design interactions with intentionally frustrated folding kinetics in order to control fuel delivery during the work cycle (24).

The present study uses efficient partition function algorithms and stochastic kinetics simulations to examine the thermodynamic and kinetic properties of sequences designed using seven methods that capture aspects of the positive and negative design paradigms. Although several of these design criteria have been widely used, we are not aware of any previous attempt to assess their relative performance. Evaluated based on thermodynamic considerations, we

*To whom correspondence should be addressed. Tel: +1 626 395 8086; Fax: +1 626 395 8845; Email: niles@caltech.edu

consistently observe that sequence selection methods that implement both positive and negative design paradigms outperform methods that implement either paradigm alone. This trend appears to be robust to changes in both the target secondary structure and the parameters in the physical model, and to the choice of either RNA or DNA as the design material. The trend does not hold when the design criteria are judged based on kinetic considerations, as favorable thermodynamic properties do not ensure fast folding.

Physical model

The secondary structure of a nucleic acid strand is simply a list of base pairs between Watson–Crick complements (A·U, C·G for RNA, and A·T, C·G for DNA) or wobble pairs (G·U, or G·T); it may be described as a graph with connections between paired bases on a polymer backbone, as depicted in Figure 1c. A coarse-grained energy landscape may be defined over the finite number of all possible secondary structures, where the properties of each secondary structure represent an ensemble average over the three-dimensional atomic structures consistent with that base-pairing graph. Decades of effort have been invested in the formulation and parameterization of an empirical potential for the free energy of a nucleic acid strand based on a loop decomposition of the base-pairing graph (25–27). Despite both conceptual and practical limitations, this model has great utility for studying the properties of natural and engineered RNA and DNA structures (10,27), serving as the basis for efficient dynamic programming algorithms that calculate the minimum energy structure (15,28–32) and partition function (21,33) for a given nucleic acid strand over a large class of secondary structures, including many pseudoknots (see Fig. 1d). [Several of these methods are now implemented for use via online servers (34,35).]

The folding kinetics of a sequence can be addressed by simulating the trajectory through secondary structure space as a continuous-time Markov process (36,37). Changes in secondary structure are described in terms of elementary steps corresponding to the breaking or formation of a single base-pair. For each elementary step, the ratio of the forward and backward rates is defined to be consistent with the equilibrium probabilities of the two end states (37,38). However, *ad hoc* arguments are required to set the magnitude of the rates. There is some evidence that qualitative properties of kinetic simulations are insensitive to the specific rate model (37).

Thermodynamic and kinetic evaluation metrics

The partition function over secondary structure space provides an ideal conceptual framework for evaluating the affinity and specificity of a sequence for the target structure. If $\Delta G(s)$ is the free energy of a sequence in secondary structure s, then the probability of sampling s at thermodynamic equilibrium is given by:

$$p(s) = \frac{1}{Q} e^{-\Delta G(s)/RT}$$

where the partition function

$$Q = \sum_{s \in \Omega} e^{-\Delta G(s)/RT}$$

is a weighted sum over the set of all secondary structures Ω, R is the universal gas constant and T is the temperature. If the probability $p(s^*)$ of folding to the target graph s^* is close to unity, then within the context of the approximate physical model, the sequence achieves both high affinity and high specificity for the target structure.

The probability $p(s^*)$ represents a very stringent design evaluation criterion since it measures the probability that every nucleotide matches the target graph exactly. For some applications (e.g. those involving large DNA molecules where some 'breathing' is unavoidable), it is acceptable to use sequences that adopt an ensemble of secondary structures similar to the target graph. In such cases, requiring $p(s^*)$ to be close to unity is a sufficient but not necessary condition for identifying satisfactory sequence designs.

A more lenient design criterion may be obtained by using a modified form of the partition function algorithm to compute the matrix of base-pair probabilities (33) with entries $P_{i,j} \in [0,1]$ corresponding to the probability of forming base pair $i{\cdot}j$. By comparing the entries of P to the structure matrix S^* with entries $S^*_{i,j} \in \{0,1\}$ describing the target secondary structure s^*, we may compute the average number of incorrect nucleotides $n(s^*)$ over the equilibrium ensemble of secondary structures Ω.

The derivation of $n(s^*)$ for a strand of length N proceeds as follows. Each secondary structure $s \in \Omega$ is defined by a symmetric $N \times N$ structure matrix S, with entries $S_{i,j} = 1$ if s contains base pair $i{\cdot}j$ and $S_{i,j} = 0$ otherwise. We augment the matrix S by adding an additional column with entries $S_{i,N+1} = 1$ if base i is unpaired and $S_{i,N+1} = 0$ otherwise. Hence, every row sum is one. Using the same convention, the augmented structure matrix corresponding to the target structure s^* is denoted S^*. Given a sequence, if the probability of sampling structure s is $p(s)$, then the average number of incorrect nucleotides may be expressed as follows:

$$n(s^*) = N - \sum_{s \in \Omega} \left[p(s) \sum_{\substack{1 \le i \le N \\ 1 \le j \le N+1}} S_{i,j} S^*_{i,j} \right].$$

This may be rearranged to give

$$n(s^*) = N - \sum_{\substack{1 \le i \le N \\ 1 \le j \le N+1}} \left[\sum_{s \in \Omega} p(s) S_{i,j} \right] S^*_{i,j},$$

where the quantity in paretheses is $P_{i,j}$, the probability of forming base pair $i{\cdot}j$. The extra column has entries $P_{i,N+1}$ equal to the probability that base i is unpaired. Again, each row sum is one. Hence the average number of incorrect nucleotides may be expressed

$$n(s^*) = N - \sum_{\substack{1 \le i \le N \\ 1 \le j \le N+1}} P_{i,j} S^*_{i,j}.$$

This metric has the advantage that sequences that adopt secondary structures similar to s^* (e.g. due to breathing) are now identified as promising candidates. However, even if $n(s^*) \ll N$, it is possible that the consistent omission or addition of certain base pairs (e.g. a hairpin stem) may cause dramatic changes to the geometric structure. The requirement that $n(s^*) \ll N$ is necessary but not sufficient to ensure that $p(s^*)$ is close to unity. On the other hand, $n(s^*) \approx 0$ is both necessary and sufficient to ensure $p(s^*) \approx 1$.

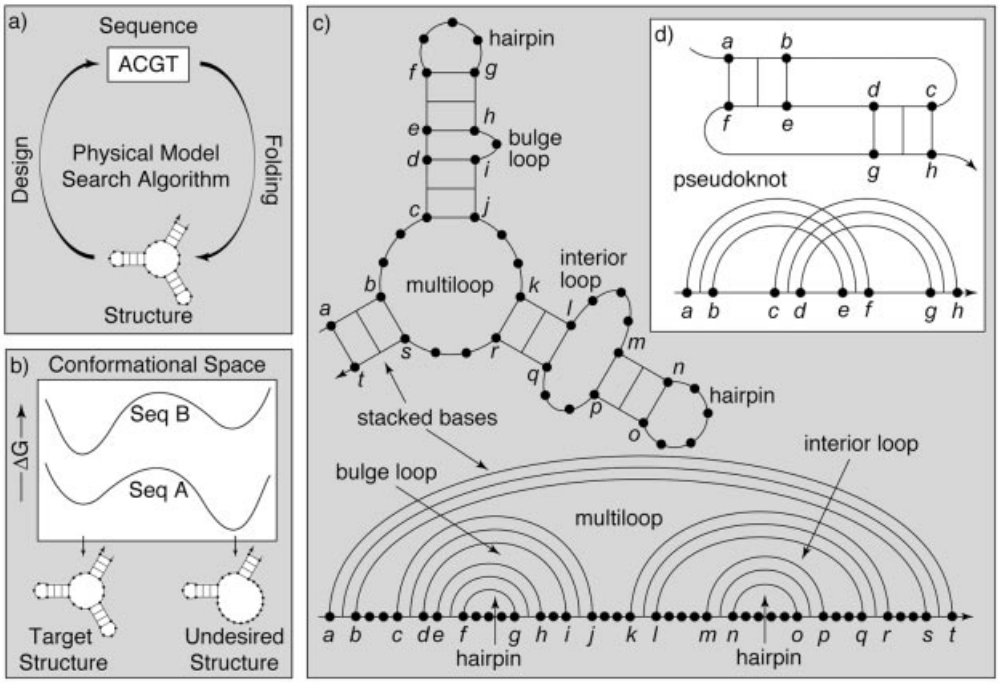

Figure 1. (**a**) Feedback loop for evaluating nucleic acid sequence designs and methodologies. (**b**) Positive and negative design paradigms. Two sequences are evaluated using an empirical potential on both the desired target structure and an undesired structure. Using a positive design paradigm, sequence A would be selected since it exhibits a stronger affinity than sequence B for the target structure (i.e. lower ΔG). Using a negative design paradigm, sequence B would be selected since it exhibits specificity for the target structure while sequence A exhibits specificity for the undesired structure. To provide a common basis for comparison, $\Delta G = 0$ for a strand with no base pairs. (**c**) Canonical loops of nucleic acid secondary structure: hairpin loops, stacked base pairs, a bulge loop, an interior loop and a multiloop. These loop structures are all nested (i.e. there are no crossing arcs in the corresponding polymer graph with the backbone drawn as a straight line). (**d**) A sample pseudoknot with base pairs $a{\cdot}f$ and $c{\cdot}h$ (with $a < c$) that fail to satisfy the nesting property $a < c < h < f$, yielding crossing arcs in the corresponding polymer graph.

We measure folding efficiency as the median time, $t(s^*)$, to achieve the target structure starting from a random coil initial condition (no secondary structure). This metric is distinct from fast folding time when the target structure is not the minimum free energy structure. Thus, $t(s^*)$ being small is neither necessary nor sufficient to imply that $p(s^*)$ is near unity. In this paper, we consider ideal sequences to be those with $p(s^*) \approx 1$, $n(s^*) \approx 0$ and $t(s^*)$ small.

Design criteria

We evaluate the following sequence design criteria:

Random. Sequences are selected to satisfy the complementarity requirements of the base-pairing graph, but are otherwise random. This is a primitive approach to both positive and negative design; compatibility with the target graph implies some affinity for the structure and incompatibility with many other graphs. At least this mild level of positive and negative design is implicit in each of the design methods that follow.

Energy minimization. Sequences are selected that attain a low energy on the target structure using the standard energy model. This approach implements explicit positive design.

Minimum free energy (MFE) satisfaction. Sequences are selected so as to ensure that the target structure is the lowest energy structure (15,19). Note that a sequence with the correct minimum energy structure may nonetheless have a low probability of adopting the target fold. This approach implements explicit negative design.

Sequence symmetry minimization (SSM). Sequences are selected so as to prohibit repeated subsequences of a specified word length (1). For subsequences that are not base-paired to consecutive bases in the target graph (e.g. single stranded or branched regions), the complementary words are also prohibited from appearing in the design. This is a heuristic approach to negative design, attempting to ensure specificity for the target structure by guaranteeing mismatches within any subsequence of the word length that hybridizes incorrectly.

Energy minimization and SSM. Sequences are selected that attain a low energy on the target graph, subject to the constraint that SSM is satisfied (14). This approach explicitly addresses both paradigms, combining rigorous positive design and heuristic negative design.

Probability. Sequences are selected to maximize the probability (15,17,21) of sampling the target structure $p(s^*)$. Positive and negative design are simultaneously addressed in a single rigorous approach.

Average incorrect nucleotides. Sequences are selected to minimize the average number of incorrect nucleotides $n(s^*)$. Positive and negative design are simultaneously addressed in a single rigorous approach.

In each case, a design method is obtained by employing a heuristic search procedure to optimize one of the design criteria. It is these design criteria that are the focus of the present work. Examining a set of sequences obtained by independent search processes provides a characterization of typical performance. For the random, MFE satisfaction and SSM methods, any sequence that satisfies the criterion is a global minimum. For the probability and average incorrect nucleotide methods, the global optimum is not necessarily attained, but there is an absolute standard of success [i.e. $p(s^*) \approx 1$ or $n(s^*) \approx 0$] that is frequently achieved. For methods involving energy minimization, there is no mathematical guarantee that the selected sequences are near the global minimum. [For energy minimization, the global minimum energy is achieved by at least one sequence for all structures we consider. For energy minimization plus SSM, we verified this property only for small structures (e.g. the one in Fig. 1a) where the global minimum was determined using a branch and bound algorithm (39) (see Algorithms).] Implementation details for all design methods are provided in Algorithms.

RESULTS

We now compare the performance of these seven design methods. All designed sequences are at local minima in the sense that no mutation of one base pair or of one unpaired base results in a better sequence based on the given design criterion.

RNA multiloop design

Each method was used to perform 100 independent sequence designs for a four-stem RNA multiloop comprising 71 nucleotides. Histograms of $p(s^*)$ and $n(s^*)$ are shown in Figure 2a and b, with median values recorded in Table 1. For random sequences, ~95% of the designs have $p(s^*) < 0.1$ and the median value of $n(s^*)$ is 7.2. Energy minimization performs worse than random while MFE satisfaction and SSM perform somewhat better. There is a dramatic improvement in sequence quality using a combination of energy minimization and SSM. Directly optimizing either $p(s^*)$ or $n(s^*)$ leads to sequences with excellent thermodynamic properties.

To provide an alternative view of average design performance, Figure 2c depicts the base-pairing probabilities $P_{i,j}$ for the median sequence based on $p(s^*)$. Energy minimization completely fails to capture the connectivity of the target structure. The other methods demonstrate the correct basic structure with varying propensities for extending or adding helices.

Model robustness

It is inevitable that new parameter sets will continue to be developed for the loop-based potential functions used for these studies (26,27). For our design methods to be useful, the quality of a sequence must be robust to perturbations in the approximate physical model; sequences that behave well using many different parameter sets are more likely to perform well in the laboratory. To examine this issue, we consider 1000 randomized potential functions for RNA where every parameter [there are 10 692 and 12 198 non-zero parameters for the RNA (27) and DNA (26) models, respectively] is independently adjusted by an amount uniformly distributed on $\pm 10\%$, $\pm 20\%$ or $\pm 50\%$.

For each design method, the top-ranked sequence based on $p(s^*)$ is re-examined using these modified potentials. The new probabilities are shown in Figure 3, with the original probabilities depicted as dashed lines. For perturbations distributed uniformly on $\pm 10\%$, these probability distributions are peaked near the original probabilities, with the sharpest peaks occurring for the best original sequences. The studies with perturbations distributed uniformly on $\pm 20\%$ and $\pm 50\%$ demonstrate that the best sequences are surprisingly robust, even to large perturbations.

Sequence composition

The contrasting behavior of sequences designed by different methods is partly attributable to the variation in sequence composition as summarized by Table 1 in terms of fraction of CG nucleotides and average Shannon entropy per position. [For 100 sequences designed by a given method, the information entropy at position i is defined by

$$\sigma_i = -\sum_{\eta = A,C,G,U} f_i(\eta) \log_4 f_i(\eta),$$

where $f_i(\eta)$ is the fraction of base η at position i, and σ_i varies between zero (all nucleotides are identical) and one (equal number of each nucleotide). The average entropy per position over a sequence of length N is then $\Sigma_{i=1,N} \sigma_i/N$.]

As expected, the random and SSM designs have a CG content of 50% and an average sequence entropy of approximately one, meaning that each base is equally likely at each position. Similar trends are observed for MFE satisfaction, emphasizing that it is a negative design approach, in that it does not attempt to optimize affinity for the target structure by increasing the CG content. In contrast, energy minimization leads to 91% CG content with a dramatic drop in the sequence entropy at each position. The combined approach of energy minimization and SSM increases the average sequence entropy and reduces the CG content to ~65%. By comparison, designs based on direct optimization of $p(s^*)$ or $n(s^*)$ have similar CG contents, but much lower average sequence entropies, suggesting greater uniformity across independent sequence designs in the placement of C and G bases throughout the strand.

The differing design objectives of methods that implement positive and negative design paradigms are amply illustrated by the plot of probability versus free energy in Figure 2d. Here, the methods of SSM and MFE satisfaction produce sequences with ΔG values comparable to those for the random method. Energy minimization naturally produces the lowest ΔG values, while the methods that combine positive and negative design sacrifice some level of affinity to achieve greater specificity and hence higher $p(s^*)$ values.

Kinetics

We estimate $t(s^*)$ as the median time to fold to s^* over 1000 stochastic simulation runs as plotted against $p(s^*)$ in Figure 2e. Each simulation was terminated after 10^4 dimensionless time units had elapsed.

Sequences designed by energy minimization had very low probabilities and failed to fold during the time frame of the

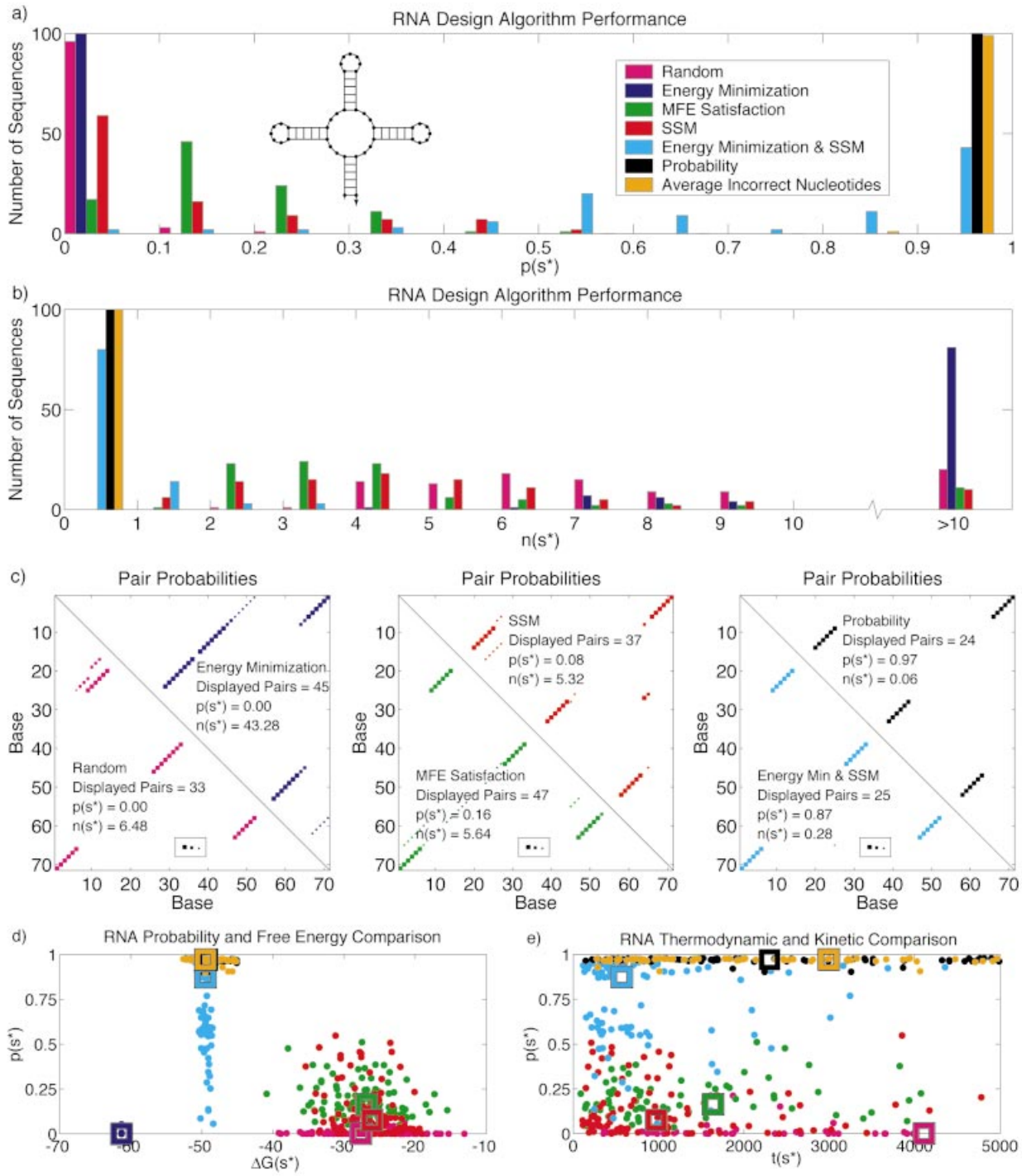

Figure 2. RNA multiloop. (**a**) Histograms for 100 sequence designs based on probability of sampling the target graph, $p(s^*)$. The color legend applies to all plots. (**b**) Histograms for the same 100 sequence designs based on average number of incorrect nucleotides, $n(s^*)$. (**c**) Base-pairing probabilities $P_{i,j}$ for the median sequence based on $p(s^*)$. Square sizes correspond to $P_{i,j} \geqslant \{0.5,0.05,0.005\}$, respectively. The target structure is identical to that obtained by optimizing probability (black) or the average number of incorrect nucleotides (not shown). (**d**) $p(s^*)$ versus free energy, $\Delta G(s^*)$. Each dot corresponds to one of 100 sequences designed using each method. Each bold square corresponds to the median over the 100 sequences designed using each method. (**e**) $p(s^*)$ versus median folding time, $t(s^*)$, over 1000 kinetic trajectories starting from random coil initial conditions. Dots and squares are interpreted as in (d).

simulations. Random sequences also had very low probabilities but did succeed in folding. On average, the negative design approaches of MFE satisfaction and SSM yielded sequences with improved probabilities and folding times relative to random sequences. The combined approach of energy minimization and SSM yielded significantly higher probabilities with folding times that are comparable to SSM. Sequences designed by direct optimization of $p(s^*)$ or $n(s^*)$ yielded the highest probabilities, but somewhat slower folding times. Figure 2e illustrates two distinct classes of slow folding

Table 1. Sequence statistics for RNA multiloop designs of Figure 2

Design method	$p(s^*)$	$n(s^*)$	CG content	Entropy	Top-ranked based on $p(s^*)$
					((((((..((((((.....))))))..((((((.....))))))..((((((.....))))))..))))))
Random	0.00	7.22	0.50	1.00	AUGGGUUAUCACUGCGGCUCAGUGAAACAAGCGUCGUUCGCUUGGGACGUCUAUAUAAGACGUUUACCCAU
Energy minimization	0.00	32.46	0.91	0.35	GGGGGCACGGGGGCCUCUGGCCCCCACGGCCCCCGCCGGGGGCCACGGGCCCCUCUGGGGCCCACGCCCCC
MFE satisfaction	0.16	4.14	0.50	0.99	GGCGUCUAAAGAACGAUAAGUUCUUAUGAUUCAAAGACUGAAUCUGGAUCGAGGACGUCGAUCGUGACGCC
SSM	0.08	4.89	0.50	0.99	GACGCACCCCUGAGACCGCCUCAGGUUGUAAGCGAUGGGCUUACCAGAUUCCACAUAGGAAUCAAUGCGUC
Energy minimization and SSM	0.87	0.28	0.66	0.68	GGAGCCAAGACCUCGUUCAGAGGUCACGCCCUGGAAAACAGGGCAACCCCGCUUAGUGCGGGGACGGCUCC
Probability	0.97	0.06	0.69	0.40	GCCGGCAAGCCCUCGACUAGAGGGCAAGCGGUCGACUAGACCGCAAGCCGUCGAAUAGACGGCAAGCCGGC
Average incorrect nucleotides	0.97	0.06	0.68	0.39	GCCGGCACGGCCUCGACUAGAGGCCAAGCCGUCGAAUAGACGGCAAGCCCUCGACUAGAGGGCAAGCCGGC

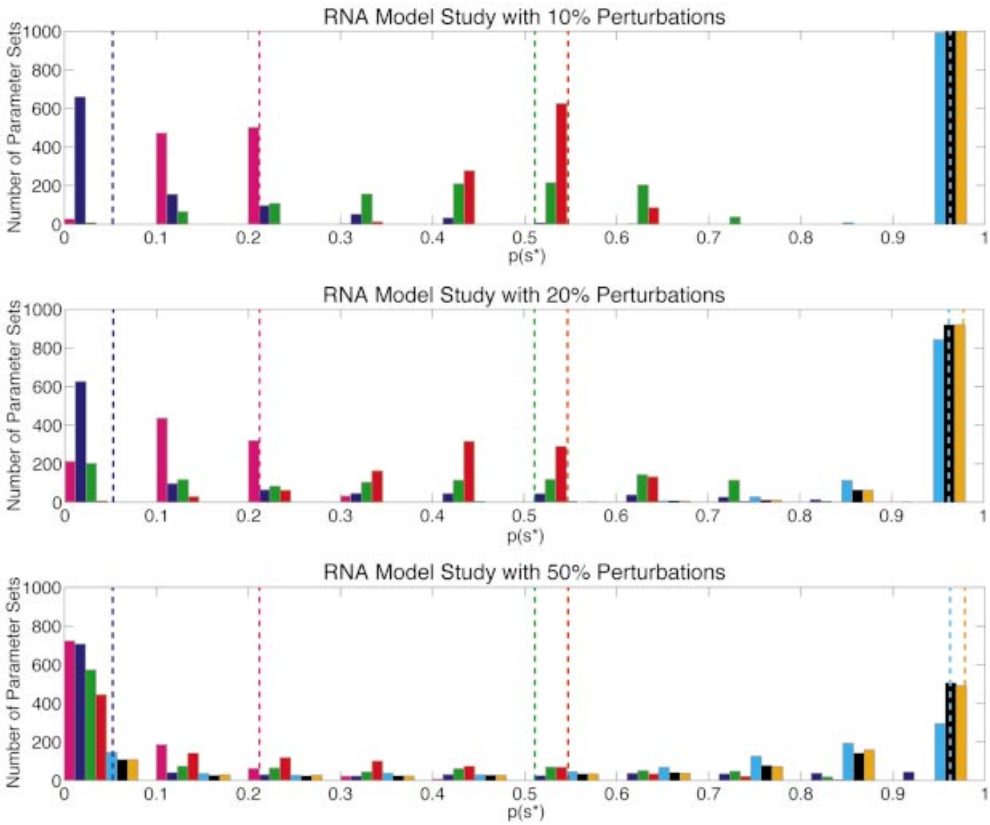

Figure 3. RNA model perturbation study. For the multiloop designs of Figure 2, the top-ranked sequence for each method based on $p(s^*)$ is re-examined using 1000 randomized potential functions where every parameter is independently adjusted by an amount uniformly distributed on ±10%, ±20% or ±50%. The original probabilities are depicted as dashed lines.

sequences: sequences with low $p(s^*)$ have energy landscapes in which s^* is not a prominent local minimum, while sequences with high $p(s^*)$ have s^* as the global minimum, but often have highly frustrated energy landscapes, possibly due to high CG content. Each of the three methods that implement both positive and negative design paradigms produce a number of sequences that appear to be excellent based on both equilibrium and kinetic properties. However, in general, the depth of the global minimum in the energy landscape does not determine the kinetic accessibility of that conformation (37).

Other RNA structures

The multiloop structure considered in Figure 2 had stems of length $\alpha = 6$ and single-stranded multiloop regions of length $\beta = 2$. In Figure 4, the design conclusions are generalized to a related family of multiloop structures with $\alpha \in \{4,6,8\}$, $\beta \in \{0,2,4\}$. Results for a larger RNA multiloop with 122 nucleotides and a small RNA pseudoknot with 30 nucleotides are shown in Figures 5 and 6. We have also examined design performance for open structures, hairpins and three-stem multiloop structures (not shown). In all of these cases, the

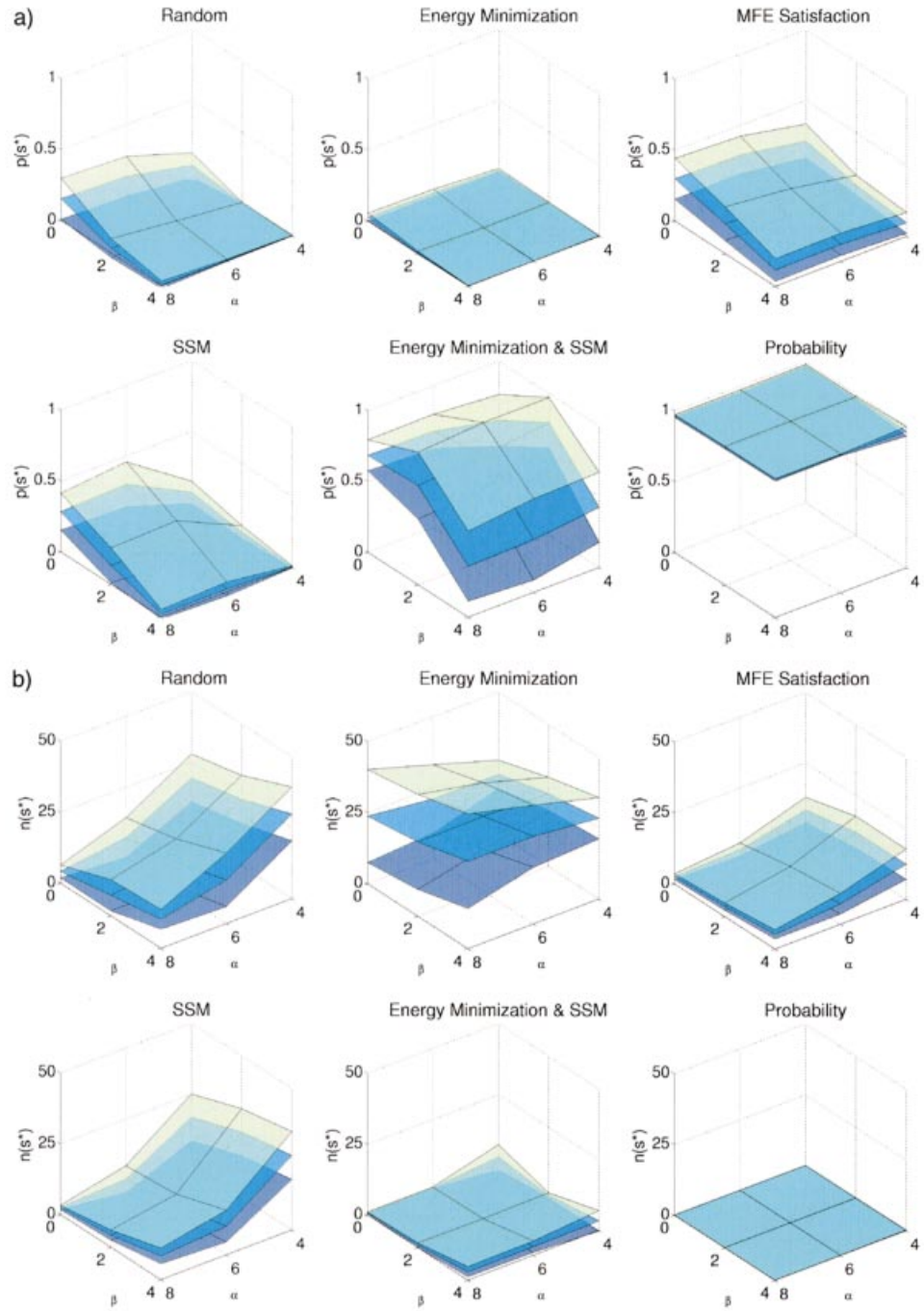

Figure 4. RNA multiloop variations. Design performance based on (**a**) $p(s^*)$ and (**b**) $n(s^*)$ with stem α = (4,6,8) and single-stranded multiloop regions β = (0,2,4). Surfaces show the mean values plus and minus one standard deviation for 100 independently designed sequences. The results for optimizing average incorrect nucleotides (not shown) are nearly indistinguishable from those obtained by optimizing probability.

same trends are observed in the relative performance of the different design methods.

DNA design

For each of the non-pseudoknotted cases, analogous data are provided for DNA in Supplementary Material (Figs 7–10 and Table 2). Similar trends are observed in the relative performance of the different design methods. Based on equilibrium properties, the most noticeable differences compared with the RNA designs are: (i) the best methods no longer consistently produce sequences with $p(s^*) > 0.90$; and (ii) structures with helices of length $\alpha = 4$ are difficult to stabilize. Comparing equilibrium and kinetic properties, higher probabilities are achievable with RNA, and faster folding times are typical for DNA.

DISCUSSION

Relative merits of design criteria

Based on thermodynamic considerations, our results support classifying design criteria according to the extent to which

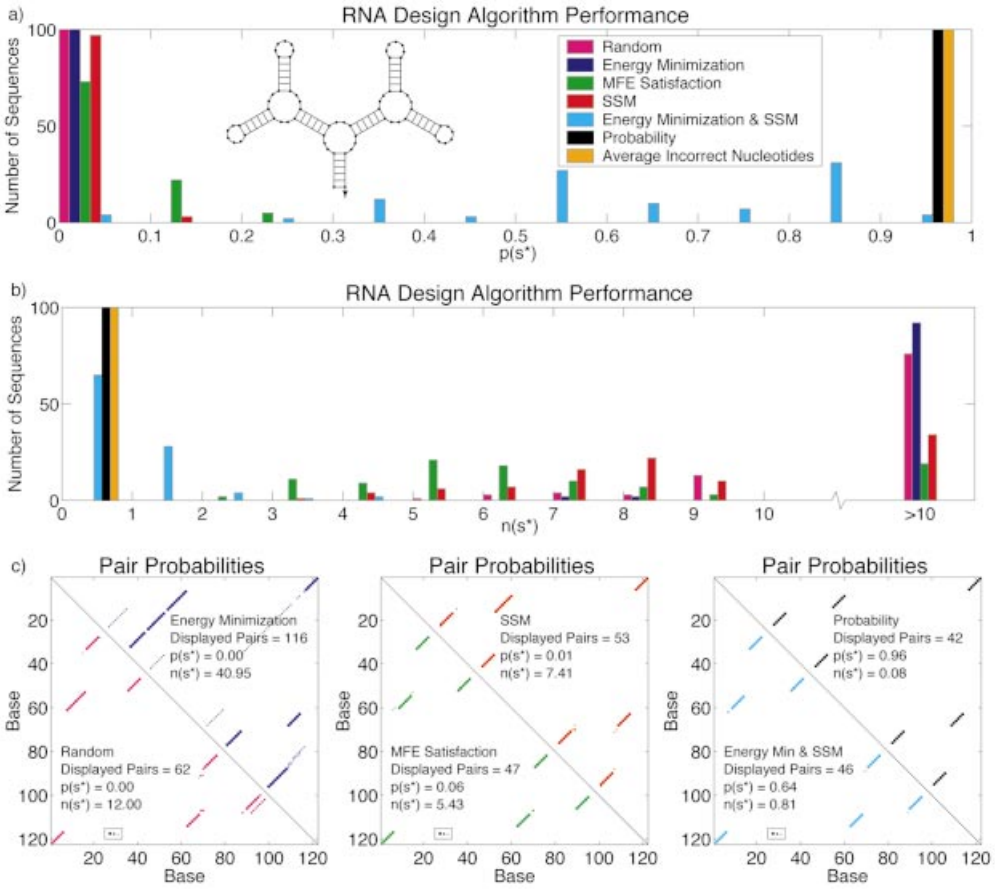

Figure 5. Large RNA multiloop. See caption for Figure 2a–c.

they implement positive (affinity) and negative (specificity) design paradigms. The design methods that implement both paradigms (energy minimization plus SSM, probability, average incorrect nucleotides) significantly outperform other methods. In general, the worst performance was observed for methods that implemented neither paradigm (random) or positive design alone (energy minimization), with somewhat better performance observed for negative design methods (MFE satisfaction, SSM). It is perhaps surprising that MFE satisfaction, which performs negative design using the full thermodynamic energy model, performs so similarly to SSM, which neglects the model. Methods based on SSM are widely used; our results suggest that they could be improved by incorporating a positive design component. For many structures, high probabilities (within the context of an approximate physical model) are obtained by directly optimizing $p(s^*)$ or $n(s^*)$.

Optimization of equilibrium properties leads to sequences with widely differing folding times. One simple design approach is to filter sequences to identify fast or slow folders as desired. Alternatively, new sequence selection algorithms could be developed that explicitly take into account the structure of the energy landscape so as to optimize the kinetic accessibility of the global minimum energy secondary structure. Furthermore, the observed decoupling of thermodynamic and kinetic properties suggests that there are sufficient degrees of freedom in sequence space to allow the design of more complex features of the energy landscape [e.g. metastable states (37)].

Robustness of claims

The consistency in the relative merits of these design methods suggests a level of generality that goes beyond the structures investigated here. It appears that it is not necessary to classify target structures according to the demands that they place on positive or negative design, as methods that implement both paradigms are generally preferred. Furthermore, we observe the same relative performance rankings for RNA and DNA despite systematically different thermodynamic parameters for the two materials. Evaluations of sequence quality for either material appear robust to perturbations in the parameter sets. This suggests that the relative merits of the design criteria are not likely to change as the empirical models are improved.

The validity of our thermodynamic metrics is linked to the validity of the underlying empirical models, which continue to be refined and evaluated by experimental studies (27,40). Further improvement of these models for both thermodynamic and kinetic predictive capability will directly benefit rational design methods. Historically, some parameters have experienced adjustments significantly larger than 10% as the model was refined (40). It seems likely that parameters that have undergone extensive study (e.g. base-pair stacking) will experience relatively small changes in the future, while other parameters that have not received the same degree of

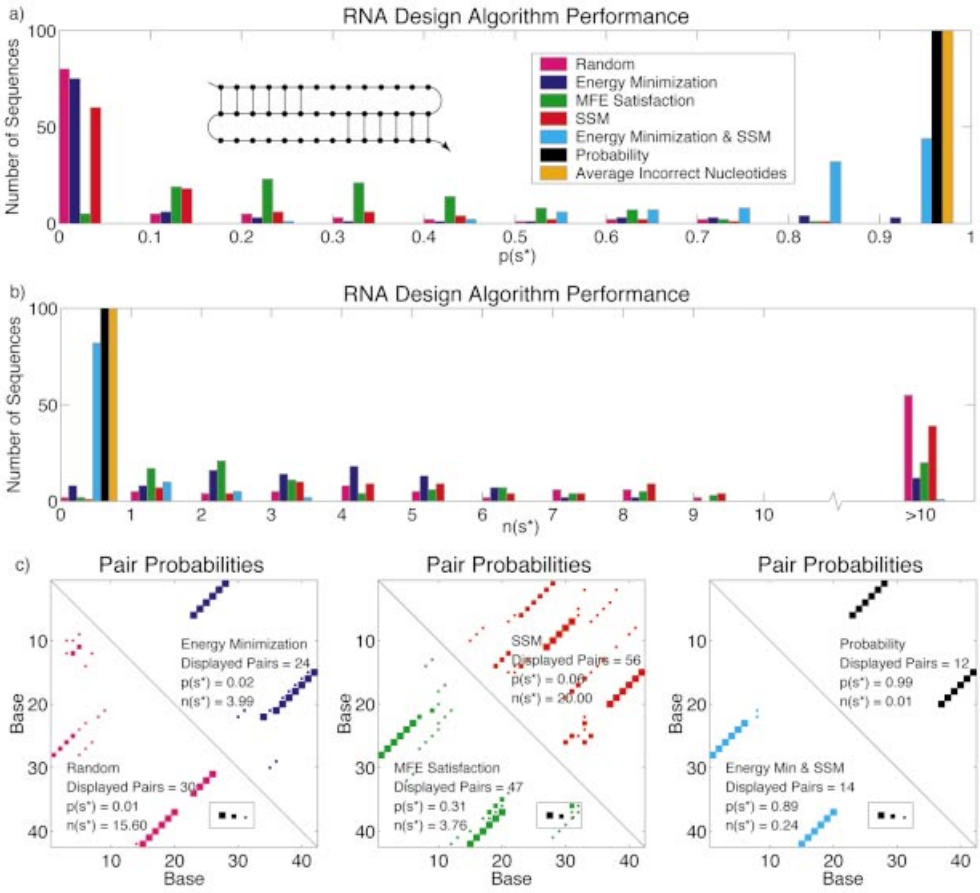

Figure 6. RNA pseudoknot. See caption for Figure 2a–c.

scrutiny (e.g. coaxial stacking or dangling ends) may change more dramatically. These adjustments could alter the design conclusions for some target structures.

The partition function may retain utility for design in certain cases where the energy model is known to be incorrect. For example, pseudoknot energy models do not fully consider geometric constraints, such as steric hindrance. Nevertheless, it is reasonable to believe that the unknown energy correction terms are non-negative; that is, structures violating geometric constraints are in fact less likely than predicted. In this case, for design targets that are geometrically unstrained (so that the missing energy term is small), the predicted $p(s^*)$ will be strictly lower than if the energy correction terms had been included (since all undesired structures have non-negative correction terms). Consequently, high-ranking sequences based on existing models are likely to be successful in practice.

Algorithmic considerations

Each design method consists of a criterion score and a heuristic for optimizing that score. Evaluating the score for a single sequence of length N is an $O(N)$ operation for random, energy minimization, SSM, and energy minimization plus SSM methods. When used in an adaptive walk, each incremental change to the score can be evaluated in constant time. For designs based on MFE satisfaction, probability, and average incorrect nucleotides, each score evaluation is an $O(N^3)$ operation if pseudoknots are excluded and an $O(N^5)$ operation if a class of pseudoknots is included. The cost of these latter methods motivates further investigation into optimization techniques for these scores, including filtering the designs of less expensive methods (14) and assembling larger structures hierarchically (15,19).

The random and SSM methods apply to single- or multi-stranded structures with or without pseudoknots. The other five methods require extensions to the standard empirical potential functions to handle multiple strands or pseudoknots. [This complication can be avoided for the two methods involving energy minimization if only stacking energies are considered and other loop terms (which are largely independent of sequence) are neglected.] The dynamic programming algorithms that underlie three of the methods also require generalization to handle problems with multiple strands (16,19,41).

Additional design constraints

Each design method considered here has been simplified to reflect the essence of the approach so as to admit easy description, comparison and replication. In practice, there are many additional considerations that might be used to modify these approaches so as to satisfy various additional constraints. For example, the designer may wish to limit the CG content, to use a three-letter alphabet (11,42), to prohibit consecutive stretches of a single base, to fix the melting temperature, or to

impose various other rules of thumb that have been garnered from years of laboratory experience. The intended function of the design may also impose additional requirements, such as the inclusion of subsequences of biological or biochemical relevance (e.g. promoters, restriction sites, genomic targets, ribozymes and deoxyribozymes). Frequently, the intention is to design a set of strands that interact to form one of several allowed secondary structures (e.g. a DNA beacon switches from a hairpin to a helix in the presence of a target ligand). In DNA computing, it is often necessary to design a combinatorial library of strands, each of which is devoid of secondary structure (18). These problems naturally lead to multi-objective optimizations, where we expect positive and negative design paradigms to continue to play a critical role.

Comparison with protein design

It is informative to compare rational nucleic acid design efforts to those in the related area of rational protein design. Proteins provide a rich design space with a much greater demonstrated range of natural function than RNA and DNA. Hence, they represent a fertile medium for the design of new medical and industrial products. While fold affinity and specificity remain fundamental design objectives for proteins, it is not clear to what degree the explicit implementation of both positive and negative design paradigms remains critical. It is possible that the biochemical properties of the 20 amino acids are sufficiently different from those of the four nucleotides that there is a change in the degree to which positive and negative design methods yield collateral specificity and affinity, respectively.

Computational models for protein thermodynamics currently require three-dimensional fold information. To stabilize a given target fold, rational design efforts have focused on positive design for fold affinity: identify the sequence with the lowest energy on the target fold (43–45). Explicit negative design for fold specificity is problematic since it is challenging to describe the space of unwanted three-dimensional folds. However, small ensembles of unwanted structures have been used to explicitly design for fold specificity (46,47). Arguments based on the random energy model suggest that implicit negative design may be achieved by fixing the sequence composition before optimization (48–50). Recently, a novel protein fold was designed from scratch (51) by alternately optimizing the sequence on a fixed backbone and the backbone for a fixed sequence. The former step represents positive design via energy minimization. The latter step was implemented by searching nearby structure space and redefining the target to be the minimum energy structure—a local form of negative design. Conceptually, it is unclear to what extent this local structural optimization implements global negative design. [The related hypothetical approach of performing global structure prediction and adjusting the target to be the MFE structure would correspond to explicit negative design (identical to MFE satisfaction except that specificity is achieved by adjusting the structure instead of the sequence).]

There is currently no physical abstraction (akin to nucleic acid secondary structure) that facilitates the prediction of protein structure from protein sequence. Hence, the feedback loop in Figure 1a must be closed either by experimental structural characterization methods or by computationally solving the protein structure prediction problem (52). This feedback can be used both to improve particular sequence designs and to improve the physical model on which the design process is based. To avoid the risk of introducing artifacts into the physical model (53), significant effort has been invested in developing exact search methods to find the globally optimal sequence based on fold affinity, although approximate search methods have also proved useful in practice (45,54).

These limitations would similarly apply to nucleic acid design based on three-dimensional atomic coordinates. However, by designing at the level of secondary structure, it is possible to address both positive and negative design paradigms explicitly and to use partition function algorithms to evaluate design quality computationally. The probability of sampling the target graph $p(s^*)$ has a maximum value of unity. Hence, it is no longer necessary to perform exact global sequence searches in order to be sure that the sequence accurately reflects the properties of the physical model; it is enough to check that $p(s^*)$ is near unity or that $n(s^*)$ is sufficiently small.

Implications for design of nanodevices

For many design applications, nucleic acids represent an attractive building material. Consider for example, an attempt to design a mechanical device that performs work by moving through a series of conformations. It would be cumbersome to parameterize the protein design problem for mechanical devices in terms of atomic coordinates. It also seems unlikely that it would be possible to conditionally stabilize a sequence of non-natural folds using positive design methods that do not explicitly treat fold specificity. However, DNA devices with moving parts and complex conditional conformational changes have already been designed (using *ad hoc* methods) and experimentally demonstrated (8,10,24). We expect that nucleic acid secondary structure will provide a productive framework for formulating the design problem for functional multi-state machines in a way that simultaneously addresses positive and negative design requirements. Ultimately, the objective of rational nucleic acid design efforts is to develop a 'molecular compiler' that takes as input a conceptual design for a device and produces, as output, a list of nucleic acid sequences that can be expected to assemble into the desired structures and function robustly.

ALGORITHMS

Design methods

The parameter sets for RNA and DNA are taken from Mfold3.1 (27), with RNA pseudoknot parameters provided by Dirks and Pierce (21). There are currently no pseudoknot parameters for DNA. Dangle energies were treated as the d2 option in the Vienna package (15). After each sequence search is performed with any of the methods described below, we check to see whether the sequence is quenched, in the sense that no mutation of a single base pair or of a single unpaired base improves the design according to the design metric. If the sequence is not quenched, we run a further adaptive walk, checking every 1000 steps to see if the sequence is quenched and terminating the search when quenching is achieved. All

RNA and DNA sequences used for these studies are provided in the Supplementary Material.

Random. One hundred random sequences are independently generated that satisfy the target graph base-pairing requirements.

Energy minimization. One hundred independent simulated annealing runs with different random initial sequences are used to identify 100 sequences with a low free energy on the target graph according to the standard loop-based energy model. Each search uses an exponentially decreasing temperature profile over 10^6 steps, where each step corresponds to a point mutation that is accepted if $\exp(-\Delta G/RT) \geqslant \rho$, where ΔG is the change in energy and $\rho \in [0,1]$ is a uniformly distributed random number.

MFE satisfaction. An adaptive walk of 1000 steps is used to identify a sequence for which the target structure is the lowest energy structure. Each step consists of a random point mutation that is accepted if the new minimum energy structure calculated using dynamic programming methods (15,21,30–32) does not increase the number of mismatches with the target graph (15,19). The 100 sequences used for the study are obtained from 100 independent searches starting from different random initial sequences. In each case, the target is the minimum energy structure.

SSM. One hundred sequences are independently selected that are compatible with the target graph and satisfy SSM (1) with word length four. For the Large Multiloop structure, the word length was increased to five to provide a larger vocabulary.

Energy minimization and SSM. A penalty term is added to the standard energy model to bias the simulated annealing search against sequences that violate SSM. The top-ranked sequences from each of 100 independent searches are free of SSM violations for the cases presented.

Probability. An adaptive walk of 1000 steps is used to search for the sequence with the highest probability of sampling the target structure based on dynamic programing calculations of the partition function (21,33). Each step consists of a random point mutation that is rejected if the probability decreases and accepted otherwise (15,17,21). The study uses the top-ranked sequence from each of 100 independent searches, starting from different random initial sequences.

Average number of incorrect nucleotides. Independent adaptive walks based on $n(s^*)$ are used to obtain 100 sequences in a manner analogous to the direct optimization of probability described above.

The formula for $n(s^*)$ is a special case of a general metric, $d(p, p')$, between two ensembles of secondary structures, p and p', that measures the average number of differing nucleotides when one secondary structure is chosen from each ensemble. By a derivation similar to the one for $n(s^*)$,

$$d(p, p') = N - \sum_{\substack{1 \leq i \leq N \\ 1 \leq j \leq N+1}} P_{i,j} P'_{i,j}.$$

The metric of Higgs and Morgan (23) is equivalent to $d(p, p)$ and $n(s^*) = d(p, s^*)$, where we abuse notation to indicate that the probability distribution is concentrated entirely on the target structure s^*.

Global energy minimization

For methods involving energy minimization, there is no mathematical guarantee that the selected sequences are near a global minimum. For small problems, the performance of heuristic search methods may be assessed by comparison to the global minimum energy obtained using an exact exponential-time branch-and-bound algorithm developed for protein design (39). If a protein is modeled as a rigid backbone with side chains represented by discrete 'rotamers', the protein design problem may be formulated as follows (43,44): given p disjoint sets of rotamers R_i (one set for each position i) and a potential function $E(\cdot,\cdot)$ that returns the energy between a pair of rotamers at different positions, choose the rotamer $r_i \in R_i$ at each position that minimizes the sum of the pairwise interaction energies between all positions:

$$E_{\text{total}} = \sum_i \sum_{j,j<i} E(r_i, r_j).$$

Methods developed for protein design may be applied to nucleic acid design if the nearest-neighbor empirical potentials (26,27) are cast as a sum of pairwise terms. For the method based on energy minimization, this is accomplished by constructing overlapping compound rotamers from nearest-neighbor bases and defining infinite energies for neighboring rotamer pairs with inconsistent overlaps. For energy minimization plus SSM, the scope of each rotamer is increased to the SSM word length and infinite energies are assigned to rotamer pairs that violate SSM.

Kinetic simulation software

Simulations are performed using Kinfold (37) with Kawasaki rate definitions based on parameter sets provided by the authors.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENTS

The authors wish to thank D. Baker, A. Condon, S. L. Mayo, N. C. Seeman and R. Schulman for comments on the manuscript, and I. Hofacker for providing the kinetic simulation package Kinfold. The following research support is gratefully acknowledged: NSF graduate research fellowship (R.M.D.), Caltech Axline SURF (M.L.), DARPA and Air Force Research Laboratory under agreement F30602-01020561 (R.M.D., E.W. and N.A.P.) and Ralph M. Parsons Foundation (N.A.P.).

REFERENCES

1. Seeman,N.C. (1982) Nucleic acid junctions and lattices. *J. Theor. Biol.*, **99**, 237–247.
2. Seeman,N.C. (1999) DNA engineering and its application to nanotechnology. *Trends Biotechnol.*, **17**, 437–443.

3. Winfree,E., Liu,F., Wenzler,L.A. and Seeman,N.C. (1998) Design and self-assembly of two-dimensional DNA crystals. *Nature*, **394**, 539–544.
4. Kallenbach,R.K., Ma,R.-I. and Seeman,N.C. (1983) An immobile nucleic acid junction constructed from oligonucleotides. *Nature*, **305**, 829–831.
5. Chen,J. and Seeman,N.C. (1991) The synthesis from DNAs of a molecule with the connectivity of a cube. *Nature*, **350**, 631–633.
6. LaBean,T.H., Yan,H., Kopatsch,J., Liu,F., Winfree,E., Reif,J.H. and Seeman,N.C. (2000) Construction, analysis, ligation and self-assembly of DNA triple crossover complexes. *J. Am. Chem. Soc.*, **122**, 1848–1869.
7. Soukup,G.A. and Breaker,R.R. (1999) Engineering precision RNA molecular switches. *Proc. Natl Acad. Sci. USA*, **96**, 3584–3589.
8. Yurke,B., Turberfield,A.J., Mills,A.P.,Jr, Simmel,F.C. and Neumann,J.L. (2000) A DNA-fuelled molecular machine made of DNA. *Nature*, **406**, 605–608.
9. Yan,H., Zhang,X., Shen,Z. and Seeman,N.C. (2002) A robust DNA mechanical device controlled by hybridization topology. *Nature*, **415**, 62–65.
10. Stojanovic,M.N. and Stefanovic,D. (2003) A deoxyribozyme-based molecular automaton. *Nat. Biotechnol.*, **21**, 1069–1074.
11. Braich,R.S., Chelyapov,N., Johnson,C., Rothemund,P.W.K. and Adleman,L. (2002) Solution of a 20-variable 3-SAT problem on a DNA computer. *Science*, **296**, 499–502.
12. Shoemaker,D.D., Lashkari,D.A., Morris,D., Mittman,M. and Davis,R.W. (1996) Quantitative phenotypic analysis of yeast deletion mutants using a highly parallel molecular bar-coding strategy. *Nature Genet.*, **16**, 450–456.
13. Brenner,S., Johnson,M., Bridgham,J., Golda,G., Lloyd,D.H., Johnson,D., Luo,S., McCurdy,S., Foy,M., Ewan,M. *et al.* (2000) Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat. Biotechnol.*, **18**, 630–634.
14. Seeman,N.C. and Kallenbach,R.K. (1983) Design of immobile nucleic acid junctions. *Biophys. J.*, **44**, 201–209.
15. Hofacker,I.L., Fontana,W., Stadler,P.F., Bonhoeffer,L.S., Tacker,M. and Schuster,P. (1994) Fast folding and comparison of RNA secondary structures. *Chem. Mon.*, **125**, 167–188.
16. Mathews,D.H., Burkard,M.E., Freier,S.M., Wyatt,J.R. and Turner,D.H. (1999) Predicting oligonucleotide affinity to nucleic acid targets. *RNA*, **5**, 1458–1469.
17. Flamm,C., Hofacker,I.L., Maurer-Stroh,S., Stadler,P.F. and Zehl,M. (2001) Design of multistable RNA molecules. *RNA*, **7**, 254–265.
18. Brenneman,A. and Condon,A. (2002) Strand design for biomolecular computation. *Theor. Comput. Sci.*, **287**, 39–58.
19. Andronescu,M., Aguirre-Hernandez,R., Condon,A. and Hoos,H.H. (2003) RNAsoft: a suite of RNA secondary structure prediction and design software tools. *Nucleic Acids Res.*, **31**, 3416–3422.
20. Kaderali,L., Deshpande,A., Nolan,J.P. and White,P.S. (2003) Primer-design for multiplexed genotyping. *Nucleic Acids Res.*, **31**, 1796–1802.
21. Dirks,R.M. and Pierce,N.A. (2003) A partition function algorithm for nucleic acid secondary structure including pseudoknots. *J. Comput. Chem.*, **24**, 1664–1677.
22. Biebricher,C.K. and Luce,R. (1992) *In vitro* recombination and terminal elongation of RNA by Qβ replicase. *EMBO J.*, **11**, 5129–5135.
23. Higgs,P.G. and Morgan,S.R. (1998) Barrier heights between ground states in a model of RNA secondary structure. *J. Phys. A: Math. Gen.*, **31**, 3153–3170.
24. Turberfield,A.J., Mitchell,J.C., Yurke,B., Mills,A.P.,Jr, Blakey,M.I. and Simmel,F.C. (2003) DNA fuel for free-running nanomachines. *Phys. Rev. Lett.*, **90**, 118102.
25. Tinoco,I.,Jr, Uhlenbec,O.C. and Levine,M.D. (1971) Estimation of secondary structure in ribonucleic acids. *Nature*, **230**, 362–367.
26. SantaLucia,J.,Jr (1996) Improved nearest-neighbor parameters for predicting DNA duplex stability. *Biochemistry*, **35**, 3555–3562.
27. Mathews,D.H., Sabina,J., Zuker,M. and Turner,D.H. (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.*, **288**, 911–940.
28. Waterman,M.S. and Smith,T.F. (1978) RNA secondary structure: a complete mathematical analysis. *Math. Biosci.*, **42**, 257–266.
29. Nussinov,R., Pieczenik,J.R., Griggs,J.R. and Kleitman,D.J. (1978) Algorithms for loop matchings. *SIAM J. Appl. Math.*, **35**, 68–82.
30. Zuker,M. and Stiegler,P. (1981) Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.*, **9**, 133–147.
31. Rivas,E. and Eddy,S.R. (1999) A dynamic programming algorithm for RNA structure prediction including pseudoknots. *J. Mol. Biol.*, **285**, 2053–2068.
32. Akutsu,T. (2000) Dynamic programming algorithms for RNA secondary structure prediction with pseudoknots. *Discrete Appl. Math.*, **104**, 45–62.
33. McCaskill,J.S. (1990) The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, **29**, 1105–1119.
34. Hofacker,I.L. (2003) Vienna RNA secondary structure server. *Nucleic Acids Res.*, **31**, 3429–3431.
35. Zuker,M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.*, **31**, 3406–3415.
36. Gillespie,D.T. (1976) General method for numerically simulating stochastic time evolution of coupled chemical reactions. *J. Comput. Phys.*, **22**, 403–434.
37. Flamm,C., Fontana,W., Hofacker,I.L. and Schuster,P. (2000) RNA folding at elementary step resolution. *RNA*, **6**, 325–338.
38. Zhang,W. and Chen,S.-J. (2002) RNA hairpin-folding kinetics. *Proc. Natl Acad. Sci. USA*, **99**, 1931–1936.
39. Gordon,D.B. and Mayo,S.L. (1999) Branch-and-terminate: a combinatorial optimization algorithm for protein design. *Structure*, **7**, 1089–1098.
40. SantaLucia,J.,Jr (1998) A unified view of polymer, dumbbell and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc. Natl Acad. Sci. USA*, **95**, 1460–1465.
41. Markham,N.R. (2003) Algorithms for nucleic acid folding, hybridization and melting prediction. Master's thesis, Rensselaer Polytechnic Institute, Troy, NY.
42. Mir,K.U. (1996) A restricted genetic alphabet for DNA computing. In Landeweber,L.F. and Baum,E.B. (eds) *DNA Based Computers II*. American Mathematical Society, Vol. 44, pp. 243–246.
43. Desjarlais,J.R. and Handel,T.M. (1995) De novo design of the hydrophobic cores of proteins. *Protein Sci.*, **4**, 2006–2018.
44. Dahiyat,B.I. and Mayo,S.L. (1997) De novo protein design: fully automated sequence selection. *Science*, **278**, 82–87.
45. Kraemer-Pecore,C.M., Wollacott,A.M. and Desjarlais,J.R. (2001) Computational protein design. *Curr. Opin. Chem. Biol.*, **5**, 690–695.
46. Havranek,J.J. and Harbury,P.B. (2003) Automated design of specificity in molecular recognition. *Nature Struct. Biol.*, **10**, 45–52.
47. Jin,W., Kambara,O., Sasakawa,H., Tamura,A. and Takada,S. (2003) De novo design of foldable proteins with smooth folding funnel: automated negative design and experimental validation. *Structure*, **11**, 581–590.
48. Shakhnovich,E.I. and Gutin,A.M. (1993) Engineering of stable and fast-folding sequences of model proteins. *Proc. Natl Acad. Sci. USA*, **90**, 7195–7199.
49. Koehl,P. and Levitt,M. (1999) De novo protein design. I. In search of stability and specificity. *J. Mol. Biol.*, **293**, 1161–1181.
50. Marshall,S.A. and Mayo,S.L. (2001) Achieving stability and conformational specificity in designed proteins via binary patterning. *J. Mol. Biol.*, **305**, 619–631.
51. Kuhlman,B., Dantas,G., Ireton,G.C., Varani,G., Stoddard,B.L. and Baker,D. (2003) Design of a novel globular protein fold with atomic-level accuracy. *Science*, **302**, 1364–1368.
52. Moult,J., Krzysztof,F., Zemla,A. and Hubbard,T. (2003) Critical assessment of methods of protein structure prediction (CASP)-round V. *Proteins*, **53**, 334–339.
53. Voigt,C.A., Gordon,D.B. and Mayo,S.L. (2000) Trading accuracy for speed: a quantitative comparison of search algorithms in protein sequence design. *J. Mol. Biol.*, **299**, 789–803.
54. Desjarlais,J.R. and Clarke,N.D. (1998) Computer search algorithms in protein modification and design. *Curr. Opin. Struct. Biol.*, **8**, 471–475.