ACS SyntheticBiology

Research Article

pubs.acs.org/synthbio

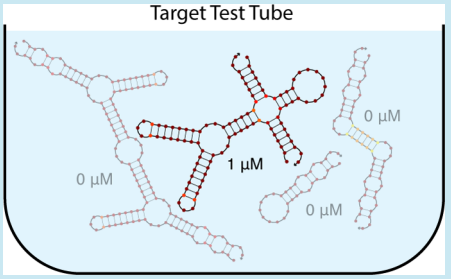
Sequence Design for a Test Tube of Interacting Nucleic Acid Strands

Brian R. Wolfe[†] and Niles A. Pierce*[,†,‡]

[†]Division of Biology and Biological Engineering and [‡]Division of Engineering and Applied Science, California Institute of Technology, Pasadena, California 91125, United States

Supporting Information

ABSTRACT: We describe an algorithm for designing the equilibrium base-pairing properties of a test tube of interacting nucleic acid strands. A target test tube is specified as a set of desired "on-target" complexes, each with a target secondary structure and target concentration, and a set of undesired "off-target" complexes, each with vanishing target concentration. Sequence design is performed by optimizing the test tube ensemble defect, corresponding to the concentration of incorrectly paired nucleotides at equilibrium evaluated over the ensemble of the test tube. To reduce the computational cost of accepting or rejecting mutations to a random initial sequence, the structural ensemble of each on-target complex is hierarchically decomposed into a tree of conditional subensembles, yielding a forest of decomposition trees. Candidate sequences are evaluated efficiently at the leaf level of the decomposition forest by estimating the test tube ensemble defect from conditional physical properties calculated over the leaf subensembles. As optimized subsequences are merged toward the root level of the forest, any emergent defects are eliminated via ensemble redecomposition and sequence reoptimization. After successfully merging subsequences to the root level, the exact test tube ensemble defect is calculated for the first time, explicitly checking for the effect of the previously neglected off-target complexes. Any off-target complexes that form at appreciable concentration are hierarchically decomposed, added to the decomposition forest, and actively destabilized during subsequent forest reoptimization. For target test tubes representative of design challenges in the molecular programming and synthetic biology communities, our test tube design algorithm typically succeeds in achieving a normalized test tube ensemble defect ≤1% at a design cost within an order of magnitude of the cost of test tube analysis.



KEYWORDS: *dilute solution, equilibrium base-pairing, target secondary structure, target concentration, test tube ensemble focusing, hierarchical ensemble decomposition, test tube ensemble defect*

The programmable chemistry of nucleic acid base pairing serves as a versatile medium for the rational design of self-assembling molecular structures, devices, and systems.[1,2] To assist in these engineering efforts, analysis algorithms have been developed to enable calculation of the equilibrium base-pairing properties of a dilute solution of interacting nucleic acid strands (e.g., a test tube), yielding predictions for the equilibrium concentration and base-pairing probabilities for an arbitrary number of complex species that form from an arbitrary number of strand species.[3−12] Of course, in an engineering setting, sequence analysis must be preceded by sequence design. However, no analogous sequence design algorithm exists for engineering the equilibrium base-pairing properties of a test tube of interacting nucleic acid strands.

To date, considerable effort has been invested in addressing the crucial subsidiary challenge of designing the equilibrium base-pairing properties of a single complex of (one or more) interacting nucleic acid strands.[5,13−29] For *complex design*, the user specifies a target secondary structure for the complex; neither the concentration of the complex, nor the concentrations of other undesired complexes are considered. As a result, sequences that are successfully optimized to stabilize a target secondary structure in the context of a complex, may nonetheless fail to ensure that this complex forms at appreciable concentration when the strands are introduced into a test tube (see Figure 1). To address this major conceptual and practical shortcoming, the present work formulates nucleic acid sequence design in the context of a test tube of interacting nucleic acid strands at equilibrium. For *test tube design*, the user specifies: (1) a set of desired "on-target" complexes, each with a target secondary structure and target concentration, (2) a set of undesired "off-target" complexes, each with vanishing target concentration.

We have previously shown that complex design can be formulated as an optimization problem based on a physically meaningful objective function, the complex ensemble defect.[12,22] For a candidate sequence and target secondary structure, the *complex ensemble defect* is the average number of incorrectly paired nucleotides at equilibrium evaluated over the ensemble of the complex.[15,22] Here, to provide a physically meaningful objective function for test tube design, we derive the *test tube ensemble defect*, corresponding to the concentration of incorrectly paired nucleotides at equilibrium evaluated over the ensemble of the test tube. To provide a basis for efficient optimization of the test tube ensemble defect, we extend

Received: March 19, 2014
Published: October 20, 2014

dx.doi.org/10.1021/sb5002196 | *ACS Synth. Biol.* 2015, 4, 1086−1100

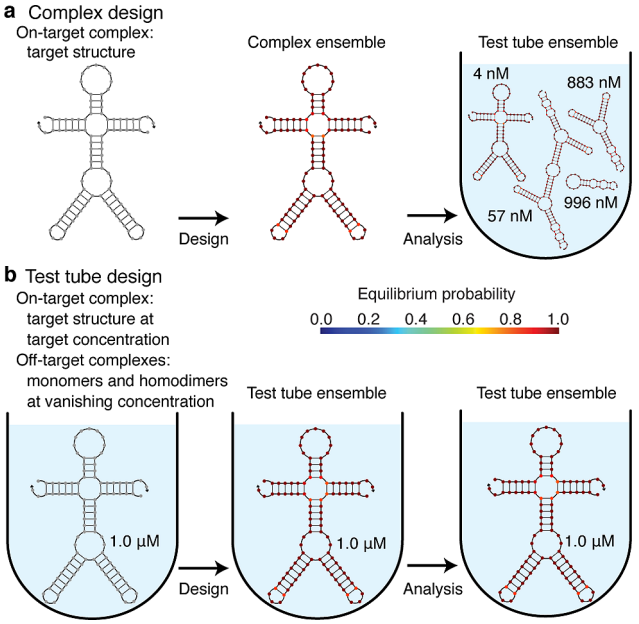


Figure 1. Complex design vs test tube design. (a) Complex design. Sequence design formulated in the context of a complex (left) ensures that at equilibrium the target structure dominates the structural ensemble of the complex (center). Unfortunately, subsequent thermodynamic analysis in the context of a test tube reveals that the desired heterodimer occurs at negligible concentration relative to other undesired monomers and homodimers (right). (b) Test tube design. Sequence design formulated in the context of a test tube (left) ensures that at equilibrium the desired "on-target" complex is dominated by its target structure and forms at approximately its target concentration and that undesired "off-target" complexes (monomers and homodimers) form at negligible concentrations (center). Subsequent thermodynamic analysis in the context of a test tube (right) is consistent with the test tube design formulation, hence providing no new information and no unpleasant surprises.

hierarchical sequence optimization concepts previously developed for complex design[5,16−18,22] and derive test tube ensemble focusing and hierarchical ensemble decomposition methods that enable efficient estimation of the test tube ensemble defect.

THEORY

We begin by describing the physical quantities that provide the basis for analyzing and designing the equilibrium base-pairing properties of a test tube of interacting nucleic acid strands.

Secondary Structure Model. The sequence, ϕ, of one or more interacting RNA strands is specified as a list of bases $\phi^a \in \{A, C, G, U\}$ for $a = 1,...,|\phi|$ (T replaces U for DNA). A *secondary structure*, s, of one or more interacting RNA strands is defined by a set of base pairs (each a Watson−Crick pair [A·U or C·G] or a wobble pair [G·U]). A *polymer graph* representation of a secondary structure is constructed by ordering the strands around a circle, drawing the backbones in succession from 5′ to 3′ around the circumference with a *nick* between each strand, and drawing straight lines connecting paired bases. A secondary structure is *unpseudoknotted* if there exists a strand ordering for which the polymer graph has no crossing lines. A secondary structure is *connected* if no subset of the strands is free of the others. A *complex* of interacting strands with strand ordering, π, has *structural ensemble*, Γ, containing all connected polymer graphs with no crossing lines.[11] (We dispense with our prior convention[11,12,22] of calling this entity an *ordered complex*). See Supporting Information section S1.3 for a discussion of distinguishability issues. For sequence ϕ and secondary structure $s \in \Gamma$, the *free energy*, $\Delta G(\phi, s)$, is calculated using nearest-neighbor empirical parameters for RNA in 1 M Na^+[30,31] or for DNA in user-specified Na^+ and Mg^{2+} concentrations.[32−34] These physical models have practical utility for the analysis[35−43] and design[44−60] of functional nucleic acid systems and provide the basis for rational analysis and design of equilibrium base-pairing in the context of a dilute solution.

Analyzing Equilibrium Base-Pairing in a Test Tube. Let Ψ^0 denote the set of strand species that interact in a test tube to form the set of complex species Ψ. For complex $j \in \Psi$, with sequence ϕ_j and structural ensemble Γ_j, the partition function

$$Q(\phi_j) = \sum_{s\in\Gamma_j} \exp[-\Delta G(\phi_j, s)/k_B T]$$

can be used to calculate the equilibrium probability of any secondary structure $s \in \Gamma_j$:

$$p(\phi_j, s) = \exp[-\Delta G(\phi_j, s)/k_B T]/Q(\phi_j)$$

Here, k_B is the Boltzmann constant and T is temperature. The equilibrium base-pairing properties of complex j are characterized by the base-pairing probability matrix $P(\phi_j)$, with entries $P^{a,b}(\phi_j) \in [0, 1]$ corresponding to the probability,

$$P^{a,b}(\phi_j) = \sum_{s\in\Gamma_j} p(\phi_j, s) S^{a,b}(s)$$

that base pair $a{\cdot}b$ forms at equilibrium within ensemble Γ_j. Here, $S(s)$ is a *structure matrix* with entries $S^{a,b}(s) = 1$ if structure s contains base pair $a{\cdot}b$ and $S^{a,b}(s) = 0$ otherwise. For convenience, the structure and probability matrices are augmented with an extra column to describe unpaired bases. The entry $S^{a,|s|+1}(s)$ is unity if base a is unpaired in structure s and zero otherwise; the entry $P^{a,|\phi_j|+1}(\phi_j) \in [0, 1]$ denotes the equilibrium probability that base a is unpaired over ensemble Γ_j. Hence, the row sums of the augmented $S(s)$ and $P(\phi_j)$ matrices are unity.

Let $Q_\Psi \equiv Q_j \; \forall j \in \Psi$ denote the set of partition functions for the complexes in the test tube. The set of equilibrium concentrations, x_Ψ, (specified as mole fractions) are the unique solution to the strictly convex optimization problem:[11]

$$\min_{x_\Psi} \sum_{j\in\Psi} x_j(\log x_j - \log Q_j - 1) \tag{1a}$$

$$\text{subject to} \quad A_{i,j}x_j = x_i^0 \quad \forall\, i \in \Psi^0 \tag{1b}$$

where the constraints impose conservation of mass. A is the stoichiometry matrix with entries $A_{i,j}$ corresponding to the number of strands of type i in complex j, and x_i^0 is the total concentration of strand i introduced to the test tube.

To analyze the equilibrium base-pairing properties of a test tube, the partition function, $Q(\phi_j)$, and equilibrium pair probability matrix, $P(\phi_j)$, must be calculated for each complex $j \in \Psi$ using $\Theta(|\phi_j|^3)$ dynamic programs.[3−11] The equilibrium concentrations, x_Ψ, are calculated by solving the convex programming problem (eq 1) using an efficient trust region method at a cost that is typically negligible by comparison.[11] The overall time complexity to analyze the test tube is then $O(|\Psi||\phi|^3_{max})$, where $|\phi|_{max}$ is the size of the largest complex.

dx.doi.org/10.1021/sb5002196 | ACS Synth. Biol. 2015, 4, 1086−1100

In specifying an analysis problem, a convenient and powerful approach is to define Ψ to include all complexes of up to L_{max} strands. For a test tube containing the set of strands, Ψ^0, the total number of complexes that can form of up to size L_{max} is[11]

$$|\Psi| = \sum_{L=1}^{L_{max}} \sum_{l=1}^{L} \frac{|\Psi^0|^{\gcd(l,L)}}{L} \tag{2}$$

leading to an overall time complexity to analyze the test tube of $O(|\Psi^0|^{L_{max}}|\phi|_{max}^3/L_{max})$.

Test Tube Design Problem Specification. A *test tube design* problem is specified as a *target test tube* containing a set of desired *on-target complexes*, Ψ^{on}, and a set of undesired *off-target complexes*, Ψ^{off}. The set of complexes in the test tube is then:

$$\Psi = \Psi^{on} \cup \Psi^{off}$$

Each complex, $j \in \Psi$, is specified as a strand ordering, π_j, corresponding to structural ensemble Γ_j. For each on-target complex, $j \in \Psi^{on}$, the user specifies a target secondary structure, s_j, and a target concentration, y_j. For each off-target complex, $j \in \Psi^{off}$, the target concentration is vanishing ($y_j = 0$) and there is no target structure ($s_j = \varnothing$). When specifying the off-targets in Ψ^{off}, it is convenient to include all complexes of up to L_{max} strands. For example, by eq 2, four strands can interact to form 108 complexes of up to size $L_{max} = 4$.

Complementarity constraints may be imposed on the design at the sequence level by defining strands in terms of *sequence domains* (e.g., see the sequence domains in the monomer and dimer on-target structures of Figure 9a) and at the structural level by specifying base-pairing within the on-target structures. Complementarity constraints can propagate between complexes if, for example, nucleotides a are b are paired in one on-target structure and nucleotides b and c are paired in another on-target structure.

Test Tube Ensemble Defect Objective Function. We seek to perform sequence optimization for test tube design based on a physically meaningful objective function that quantifies sequence quality with respect to the target test tube.

As a precedent for this approach, consider the related problem of complex design, where the goal is to design strands that, at equilibrium, adopt a target secondary structure within the ensemble of a complex. For a candidate sequence, ϕ_j, and target structure, s_j, the complex ensemble defect[15,22]

$$n(\phi_j, s_j) = |\phi_j| - \sum_{\substack{1 \le a \le |\phi_j| \\ 1 \le b \le |\phi_j|+1}} P^{a,b}(\phi_j) S(s_j)$$

is the average number of incorrectly paired nucleotides at equilibrium evaluated over the ensemble of the complex, Γ_j. The complex ensemble defect falls in the interval $(0, |\phi_j|)$. For complex design, the complex ensemble defect provides a physically meaningful objective function for quantifying sequence quality.

Here, to provide a basis for test tube design, we derive the test tube ensemble defect, representing the concentration of incorrectly paired nucleotides at equilibrium evaluated over the ensemble of the test tube. For candidate sequences, ϕ_Ψ, and a target test tube with target secondary structures, s_Ψ, and target concentrations, y_Ψ, the test tube ensemble defect

$$C(\phi_\Psi, s_\Psi, y_\Psi) = \sum_{j \in \Psi} c(\phi_j, s_j, y_j) \tag{3}$$

may be expressed in terms of the defect contribution of each complex $j \in \Psi$:

$$c(\phi_j, s_j, y_j) = n(\phi_j, s_j)\min(x_j, y_j) + |\phi_j| \max(y_j - x_j, 0) \tag{4}$$

For each on-target complex, $j \in \Psi^{on}$, the first term in eq 4 represents the *structural defect*, quantifying the concentration of nucleotides that are in an incorrect base-pairing state on average within the ensemble of complex j, and the second term represents the *concentration defect*, quantifying the concentration of nucleotides that are in an incorrect base-pairing state because there is a deficiency in the concentration of complex j. Because $y_j = 0$ for off-target complexes, the structural and concentration defects are both identically zero (so the sum in eq 3 may be written over Ψ^{on} instead of Ψ). This does not mean that the defects associated with the off-targets are ignored. By conservation of mass, nonzero off-target concentrations imply deficiencies in on-target concentrations, and these concentration defects are quantified by eq 3. The test tube ensemble defect falls in the interval $(0, y_{nt})$, where

$$y_{nt} \equiv \sum_{j \in \Psi^{on}} |\phi_j| y_j$$

is the total concentration of nucleotides in the test tube.

Note that if there is only one species of complex in the test tube ($|\Psi| = 1$), its concentration is necessarily equal to the target concentration ($x_1 = y_1$), so the formulation is independent of concentration. In this case, optimization of the test tube ensemble defect, $C(\phi_1, s_1, y_1)$, is equivalent to optimization of the complex ensemble defect, $n(\phi_1, s_1)$.

Calculation of the test tube ensemble defect (eq 3) requires calculation of the complex partition functions, Q_Ψ, which are used to calculate the equilibrium concentrations, x_Ψ, as well as the equilibrium pair probability matrices, $P_{\Psi^{on}}$, which are used to calculate the complex ensemble defects, $n_{\Psi^{on}}$. Hence, the time complexity to evaluate the test tube ensemble defect is the same as the time complexity to analyze equilibrium base-pairing in a test tube.

■ ALGORITHM

Overview. We describe a test tube design algorithm based on test tube ensemble defect optimization. For a target test tube with target secondary structures, s_Ψ, and target concentrations, y_Ψ, we seek to design a set of sequences, ϕ_Ψ, such that the test tube ensemble defect satisfies the *test tube stop condition*:

$$C(\phi_\Psi, s_\Psi, y_\Psi) \le C_{stop} \tag{5}$$

with

$$C_{stop} \equiv f_{stop} y_{nt}$$

for a user-specified value of $f_{stop} \in (0, 1)$.

The test tube ensemble defect is reduced via iterative mutation of a random initial sequence. Because of the high computational cost of calculating the test tube ensemble defect, it is important to avoid direct recalculation of C in evaluating each candidate mutation. We exploit two approximations to enable efficient estimation of the test tube ensemble defect: using *test tube ensemble focusing*, sequence optimization initially

focuses on only the on-target portion of the test tube ensemble; using *hierarchical ensemble decomposition*, the structural ensemble of each on-target complex is hierarchically decomposed into a tree of conditional subensembles, yielding a forest of decomposition trees. Candidate sequences are evaluated efficiently by estimating the test tube ensemble defect from conditional physical properties calculated over the conditional structural ensembles at the leaf level of the decomposition forest. As optimized subsequences are merged toward the root level of the forest, any emergent defects are eliminated via ensemble redecomposition from the parent level on down and sequence reoptimization from the leaf level on up. After subsequences are successfully merged to the root level, the exact test tube ensemble defect, C, is calculated for the first time, explicitly checking for the effect of the previously neglected off-target complexes. Any off-target complexes observed to form at appreciable concentration are hierarchically decomposed, added to the decomposition forest, and actively destabilized during subsequent forest reoptimization. The elements of this hierarchical sequence optimization algorithm are described below and in the pseudocode of Supporting Information Algorithm S1.

Test Tube Ensemble Focusing. To reduce the cost of sequence optimization, the set of complexes, Ψ, is partitioned into two disjoint sets:

$$\Psi = \Psi^{\text{active}} \cup \Psi^{\text{passive}}$$

where Ψ^{active} denotes complexes that will be actively designed and Ψ^{passive} denotes complexes that will inherit sequence information from Ψ^{active}. Only the complexes in Ψ^{active} are directly accounted for in the focused test tube ensemble that is used to evaluate candidate sequences. Initially, we set

$$\Psi^{\text{active}} = \Psi^{\text{on}}, \quad \Psi^{\text{passive}} = \Psi^{\text{off}}$$

so that only the on-target complexes are included in the focused test tube ensemble at the outset of sequence design.

Hierarchical Ensemble Decomposition. Exact evaluation of the test tube ensemble defect, C, requires calculation of the defect contribution, c_j, for each complex $j \in \Psi$. The $\Theta(|\phi_j|^3)$ cost of calculating c_j is dominated by calculation of the partition function, Q_j, and equilibrium pair probability matrix, P_j. To reduce the cost of evaluating candidate sequences, we seek to estimate c_j at lower cost by hierarchically decomposing the structural ensemble, Γ_j, of each complex $j \in \Psi^{\text{active}}$ into a tree of subensembles, yielding a forest of $|\Psi^{\text{on}}|$ decomposition trees. Each node in the forest is indexed by a unique integer k. Estimating the defect contribution, c_j, using physical quantities calculated at depth d requires calculation of the nodal partition function, Q_k, and nodal pair probability matrix, P_k, at cost $\Theta(|\phi_k|^3)$ for each node k at depth d in the decomposition tree of complex j. For an optimal binary decomposition, $|\phi_k|$ halves and the number of nodes doubles at each depth moving down the tree, so the cost of estimating c_j at depth d can be a factor of $1/2^{2d-2}$ lower than the cost of calculating c_j exactly on the full ensemble Γ_j. Hence, for maximum efficiency, candidate mutations are evaluated based on the estimated test tube ensemble defect calculated at the leaves of the decomposition forest. As designed subsequences are merged toward the root level, the test tube ensemble defect is estimated at intermediate depths in the forest.

To decompose the structural ensemble Γ_k of parent node k, the nucleotides of k are partitioned into left and right child nodes k_l and k_r by a *split-point* F (Figure 2a). In each child ensemble, the base pair adjacent to F is enforced, leading to conditional child ensembles, $\tilde{\Gamma}_{k_l}$ and $\tilde{\Gamma}_{k_r}$, that can be used to reconstruct the conditional parent ensemble, $\tilde{\Gamma}_k$, which contains all structures in Γ_k that contain the two base pairs that sandwich F. For the purposes of accuracy, it is important that $\tilde{\Gamma}_k$ should include those structures that dominate the equilibrium physical properties of Γ_k, while, for the purposes of efficiency, it is important that $\tilde{\Gamma}_k$ should exclude as many structures as possible that contribute negligibly to the equilibrium physical properties of Γ_k. Hence, the utility of ensemble decomposition hinges on suitable placement of the split-point F within parent node k.

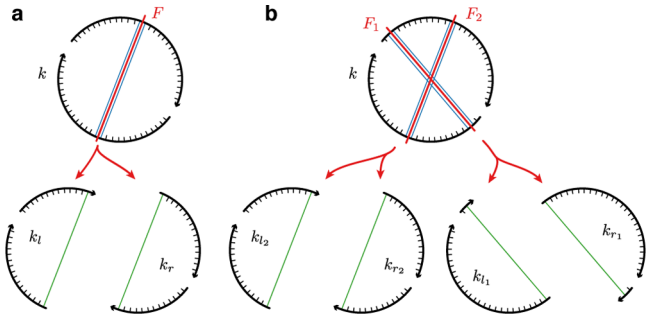


Figure 2. Ensemble decomposition of a parent node using one or more split-points sandwiched by base pairs. (a) The single split-point, F, partitions the nucleotides of parent node, k, into left and right child nodes k_l and k_r. (b) The two split-points, F_1 and F_2, cross in the polymer graph (denoted $F_1 \otimes F_2$) and are hence exclusive. Split-points within each parent are depicted in red, sandwiching base pairs within each parent are depicted in blue, and base pairs that are enforced within each child ensemble are depicted in green.

The dual goals of accuracy and efficiency can both be achieved by placing the split-point F within a duplex that forms with high probability at equilibrium such that approximately half the parent nucleotides are partitioned to each child. Recall that the structural ensemble Γ_k is defined to contain all unpseudoknotted secondary structures, corresponding to precisely those polymer graphs with no crossing base pairs. Since no structure in Γ_k can contain both base pairs that sandwich F and base pairs that cross F, placement of F between base pairs with probability close to one implies that the structures containing base pairs crossing F occur with low probability at equilibrium and may be safely neglected. Partitioning the parent nucleotides into left and right children of equal size minimizes the total cost, $\Theta(|\phi_{k_l}|^3) + \Theta(|\phi_{k_r}|^3)$, of evaluating both children.

During the course of sequence design, if the base pairs sandwiching split-point F in parent k do not form with probability close to one, the accuracy of the decomposition breaks down. In this case, $\tilde{\Gamma}_k$ excludes structures that are important to the equilibrium physical properties of Γ_k, preventing the children from approximating the full defect of the parent. As we describe later, the resulting decomposition defects are eliminated by redecomposing the parental ensemble, Γ_k, using a set of multiple exclusive split-points, $\{F\}$, that define exclusive child subensembles (Figure 2b), again enabling accurate estimation of the parental physical properties.

Structure-Guided Decomposition of On-Target Complexes. At the outset of sequence design, equilibrium base-pairing probabilities are not yet available to guide ensemble decomposition. Instead, initial decomposition of each on-target complex, $j \in \Psi^{\text{active}}$, is guided by the user-specified on-target structure, s_j, making the optimistic assumption that the base

dx.doi.org/10.1021/sb5002196 | ACS Synth. Biol. 2015, 4, 1086−1100

pairs in s_j will form with probability close to one after sequence design. Using this structure-guided ensemble decomposition approach, as the quality of the sequence design improves, the quality of the ensemble decomposition approximation will also improve.

For each complex $j \in \Psi^{\text{active}}$, the target structure s_j is decomposed into a (possibly unbalanced) binary tree of substructures, resulting in a forest of $|\Psi^{\text{on}}|$ trees. Each nucleotide in parent structure s_k is partitioned to either the left or right child substructure ($s_k = s_{k_l} \cup s_{k_r}$, $s_{k_l} \cap s_{k_r} = \varnothing$) via decomposition at a split-point F between base pairs within a duplex stem of s_k. Eligible *split-points* are those locations within a duplex stem with at least H_{split} consecutive base pairs on either side, such that each child would have at least N_{split} nucleotides. An eligible split-point is selected so as to minimize the difference in the number of nucleotides in each child, $||\phi_{k_l}| - |\phi_{k_r}||$. See Figure 3 for an example of structure-guided hierarchical ensemble decomposition.

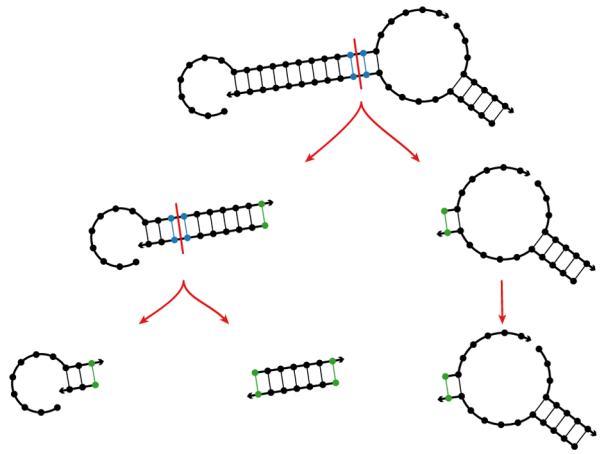


Figure 3. Structure-guided hierarchical ensemble decomposition of an on-target complex. Split-points within each parent are depicted in red, sandwiching base pairs within each parent are depicted in blue, and base pairs that are enforced within each child ensemble are depicted in green.

If the maximum depth of a leaf in the forest of binary trees is D, any nodes with depth $d < D$ that lack an eligible split-point are replicated at each depth down to D so that all leaves have depth D. Let Λ denote the set of all nodes in the forest. Let Λ_d denote the set of all nodes at depth d.

Stop Condition Stringency. In order to build in tolerance for a basal level of decomposition defect as subsequences are merged moving up the decomposition forest, the stringency of the test tube stop condition (eq 5) is increased by a factor of $f_{\text{stringent}} \in (0, 1)$ at each level moving down the decomposition forest:

$$C_d^{\text{stop}} \equiv C_{\text{stop}}(f_{\text{stringent}})^{d-1} \quad \forall\, d \in \{1, \ldots, D\}$$

Efficient Estimation of Test Tube Ensemble Properties. In the following sections, we describe how to calculate physical quantities at any level $d \in \{2,\ldots,D\}$ so as to efficiently and accurately estimate the complex contributions, $c_{\Psi^{\text{active}}}$, to the test tube ensemble defect, C. The complex partition function estimates, $\tilde{Q}_{\Psi^{\text{active}}}$, are constructed from the conditional partition functions, $\tilde{Q}_{\Lambda_d}$, and the complex pair probability matrix estimates, $\tilde{P}_{\Psi^{\text{active}}}$, are constructed from the conditional pair probability matrices, $\tilde{P}_{\Lambda_d}$. Complex concentration estimates, $\tilde{x}_{\Psi^{\text{active}}}$, are then calculated based on $\tilde{Q}_{\Psi^{\text{active}}}$, using deflated mass constraints to model the effect of the neglected off-target complexes in Ψ^{passive}. Complex ensemble defect estimates, $\tilde{n}_{\Psi^{\text{on}}}$, are calculated based on $\tilde{P}_{\Psi^{\text{on}}}$. These estimates are then used to calculate the defect estimates, $\tilde{c}_{\Psi^{\text{on}}}$, which are summed to produce the test tube ensemble defect estimate, $\tilde{C}_d$.

Complex Partition Function Estimate. We begin by calculating the complex partition function estimate, $\tilde{Q}_j$, for each complex $j \in \Psi^{\text{active}}$ in terms of conditional partition functions evaluated efficiently at any depth $d \in \{2,\ldots,D\}$.

Consider node k at depth d. Let E_k denote the set of base pairs that are enforced in node k and are hence adjacent to a split-point in some ancestor. On node k, we calculate the conditional partition function

$$\tilde{Q}_k \equiv Q(\phi_k|E_k) \tag{6}$$

over the conditional ensemble, $\tilde{\Gamma}_k$, comprising all structures in Γ_k that contain all base pairs in E_k. This calculation is performed using a dynamic program suitable for complexes containing arbitrary numbers of strands,[11] enforcing the base pairs in E_k by applying a bonus energy, ΔG^{clamp}, each time a base pair in E_k is encountered within the dynamic program (see Supporting Information section S1.2).[31,61]

Next, the conditional partition functions calculated at depth d are merged recursively to estimate the partition function for complex j. Consider split-point F in parent k, with left-child and right-child conditional partition functions, $\tilde{Q}_{k_l}$ and $\tilde{Q}_{k_r}$, and free energy, $\Delta G_F^{\text{interior}}$, for the interior loop formed by the base pairs sandwiching F (see Supporting Information section S1.2). The partition function estimate for parent k is then

$$\tilde{Q}_k = \tilde{Q}_{k_l}\tilde{Q}_{k_r}\exp(-\Delta G_F^{\text{interior}}/k_B T) \tag{7}$$

At the conclusion of recursive merging, the partition function estimate for complex j based on conditional quantities calculated at depth d is $\tilde{Q}_j$. This estimate becomes exact as the equilibrium probabilities of the base pairs sandwiching decomposition split-points approach unity in accordance with the decomposition assumption. In this case, enforcing these base pairs within the conditional child ensembles at depth d leads to conditional child partition function estimates neglecting only structures that contribute negligibly to the partition function of complex j.

Complex Pair Probability Matrix Estimate. Similarly, on node k, we calculate the conditional pair probability matrix

$$\tilde{P}_k \equiv P(\phi_k|E_k) \tag{8}$$

over the conditional ensemble, $\tilde{\Gamma}_k$, using a related dynamic program.[11] The conditional pair probability matrices calculated at depth d are merged recursively to calculate the pair probability matrix estimate for complex j. During each merge, the matrix entries for the parent are taken from the corresponding entries in the children. At the conclusion of recursive merging, the pair probability matrix estimate for complex j based on conditional quantities calculated at depth d is $\tilde{P}_j$. This estimate becomes exact in the limit as the equilibrium probabilities of the base pairs sandwiching the decomposition split-points approach unity in accordance with the decomposition assumption. In this case, enforcing these base pairs within the conditional child ensembles at depth d is an accurate

reflection of the predominant equilibrium base-pairing state of these nucleotides in complex j.

Complex Concentration Estimate using Deflated Mass Constraints. After calculating the set of complex partition function estimates, $\tilde{Q}_{\Psi^{\text{active}}}$, based on the conditional partition functions at level d, the corresponding equilibrium complex concentration estimates, $\tilde{x}_{\Psi^{\text{active}}}$, may be found by solving the convex programming problem (eq 1). To impose the conservation of mass constraints (eq 1b), the total concentration of each strand species, $i \in \Psi^0$, must be specified. The total strand concentrations

$$x_i^0 = \sum_{j \in \Psi^{\text{on}}} A_{i,j} y_j \quad \forall\, i \in \Psi^0 \tag{9}$$

follow from the target concentration, y_j, and strand composition, $A_{i,j}$, of each on-target complex $j \in \Psi^{\text{on}}$.

Using test tube ensemble focusing, initial sequence optimization is performed on a decomposition forest that contains only the on-target complexes in Ψ^{active}, but ultimately, we wish to satisfy the test tube stop condition (eq 5) for the full set of complexes in Ψ, including the off-targets in Ψ^{passive}. Recall that the off-targets in Ψ^{passive} do not contribute directly to the sum used to calculate the test tube ensemble defect (eq 3) but contribute indirectly by forming with positive concentrations, causing concentration defects for complexes in Ψ^{active} as a result of conservation of mass. Hence, we can preallocate a portion of the permitted test tube ensemble defect, $f_{\text{stop}} y_{\text{nt}}$, to the neglected off-target complexes in Ψ^{passive} by deflating the total strand concentrations (eq 9) used to impose the mass constraints (eq 1b) in calculating the equilibrium concentrations $\tilde{x}_{\Psi^{\text{active}}}$.

Following this approach, if $\Psi^{\text{passive}} \neq \varnothing$, we make the assumption that the complexes in Ψ^{passive} consume a constant fraction of each total strand concentration:

$$\sum_{j \in \Psi^{\text{passive}}} A_{i,j} \tilde{x}_j = f_{\text{passive}} f_{\text{stop}} \sum_{j \in \Psi^{\text{on}}} A_{i,j} y_j \quad \forall\, i \in \Psi^0$$

with $f_{\text{passive}} \in (0,1)$, corresponding to a total mass allocation of $f_{\text{passive}} f_{\text{stop}} y_{\text{nt}}$ to the neglected off-targets in Ψ^{passive}. To calculate the complex concentration estimates, $\tilde{x}_{\Psi^{\text{active}}}$, via eq 1, we therefore use the deflated strand concentrations:

$$x_i^0 = (1 - f_{\text{passive}} f_{\text{stop}}) \sum_{j \in \Psi^{\text{on}}} A_{i,j} y_j \quad \forall\, i \in \Psi^0 \tag{10}$$

in place of the full strand concentrations (eq 9). In the context of the deflated-mass test tube, the complex concentration estimates, $\tilde{x}_{\Psi^{\text{active}}}$, become exact in the limit as the complex partition function estimates, $\tilde{Q}_{\Psi^{\text{active}}}$, become exact.

Complex Ensemble Defect Estimate. For each complex $j \in \Psi^{\text{on}}$, the complex ensemble defect estimate, $\tilde{n}_j$, is calculated using the complex pair probability matrix estimate, $\tilde{P}_j$, reconstructed from conditional quantities calculated efficiently at any depth $d \in \{2,\ldots,D\}$.

For complex j, the contribution of nucleotide a to the complex ensemble defect estimate is given by

$$\tilde{n}_j^a = 1 - \sum_{1 \le b \le |\phi_j|+1} \tilde{P}_j^{a,b} S_j^{a,b}$$

and the complex ensemble defect estimate is then

$$\tilde{n}_j = \sum_{1 \le a \le |\phi_j|} \tilde{n}_j^a$$

This estimate becomes exact in the limit as the complex pair probability matrix estimate, $\tilde{P}_j$, becomes exact.

Test Tube Ensemble Defect Estimate. Having calculated the complex concentration estimates, $\tilde{x}_{\Psi^{\text{active}}}$, and the complex ensemble defect estimates, $\tilde{n}_{\Psi^{\text{on}}}$, based on conditional quantities calculated efficiently at any depth $d \in \{2,\ldots,D\}$, the test tube ensemble defect estimate is

$$\tilde{C}_d = \sum_{j \in \Psi^{\text{on}}} \tilde{c}_j \tag{11}$$

where

$$\tilde{c}_j = \tilde{n}_j \min(\tilde{x}_j, y_j) + |\phi_j| \max(y_j - \tilde{x}_j, 0)$$

is the contribution of complex j. In the context of the deflated-mass test tube, this estimate becomes exact in the limit as the concentration estimates, $\tilde{x}_{\Psi^{\text{active}}}$, and complex ensemble defect estimates, $\tilde{n}_{\Psi^{\text{on}}}$, become exact.

Finally, we note that the contribution of nucleotide a in complex $j \in \Psi^{\text{on}}$ to the test tube ensemble defect estimate, $\tilde{C}_d$, is

$$\tilde{C}_d^a = \tilde{n}_j^a \min(\tilde{x}_j, y_j) + \max(y_j - \tilde{x}_j, 0)$$

which will provide a basis for defect-weighted mutation sampling during leaf mutation and defect-weighted reseeding during leaf reoptimization.

Sequence Optimization at the Leaves of the Decomposition Forest. *Initialization.* The sequences for the complexes $j \in \Psi^{\text{active}}$ are randomly initialized subject to complementarity constraints in the design problem specification: Watson–Crick complements are used to initialize complementary sequence domains or any bases that are paired within an on-target structure. These initial sequences are pushed down to the leaf level of the decomposition forest.

Leaf Mutation. To minimize computational cost, all candidate mutations are evaluated at the leaf nodes, $k \in \Lambda_D$, of the decomposition forest. Leaf mutation terminates successfully if the *leaf stop condition*,

$$\tilde{C}_D \le C_D^{\text{stop}} \tag{12}$$

is satisfied. A candidate mutation is accepted if it decreases the test tube ensemble defect estimate (eq 11) and rejected otherwise.

We perform *defect-weighted mutation sampling* by selecting nucleotide a for mutation with probability $\tilde{C}_D^a / \tilde{C}_D$, proportional to the contribution of nucleotide a to the test tube ensemble defect estimate. If the selected candidate mutation position is subject to complementarity constraints implied by the design problem specification, either via complementary sequence domains or via base-pairing within an on-target structure, the candidate mutation respects the constraint with either Watson–Crick complementarity (default; constrained nucleotides are selected randomly from a uniform distribution of Watson–Crick pairs) or wobble complementarity (constrained nucleotides are selected randomly from a uniform distribution of Watson–Crick and wobble pairs). For design problems where on-target structures place competing demands on the test tube ensemble defect, permitting wobble complementarity gives the algorithm additional flexibility in meeting these demands (e.g., see the example of Figure 9).

dx.doi.org/10.1021/sb5002196 | ACS Synth. Biol. 2015, 4, 1086–1100

A candidate sequence, $\hat{\phi}_{\Lambda_D}$, is evaluated via calculation of the test tube ensemble defect estimate, $\tilde{C}_D$, if the candidate mutation, ξ, is not in the set of previously rejected mutations, γ_{bad} (position and sequence). The set, γ_{bad}, is updated after each unsuccessful mutation and cleared after each successful mutation. The counter m_{bad} is used to keep track of the number of consecutive failed mutation attempts; it is incremented after each unsuccessful mutation and reset to zero after each successful mutation. Leaf mutation terminates unsuccessfully if $m_{bad} \geq M_{bad}$. The outcome of leaf mutation is the set of leaf sequences, ϕ_{Λ_D}, corresponding to the lowest encountered $\tilde{C}_D$.

Leaf Reoptimization. After leaf mutation terminates, if the leaf stop condition (eq 12) is not satisfied, leaf reoptimization commences. At the outset of each round of leaf reoptimization, we perform *defect-weighted reseeding* of M_{reseed} positions by selecting nucleotide a for reseeding (with a new random initial sequence) with probability $\tilde{C}_D^a/\tilde{C}_D$, proportional to the contribution of nucleotide a to the test tube ensemble defect estimate. Complementarity constraints are respected as during leaf mutation. After performing a new round of leaf mutation, the reoptimized candidate sequences, $\hat{\phi}_{\Lambda_D}$, are accepted if they decrease $\tilde{C}_D$ and rejected otherwise. The counter m_{reopt} is used to keep track of the number of rounds of leaf reoptimization; m_{reopt} is incremented after each rejection and reset to zero after each acceptance. Leaf reoptimization terminates successfully if the leaf stop condition is satisfied and unsuccessfully if $m_{reopt} \geq M_{reopt}$. The outcome of leaf reoptimization is the set of leaf sequences, ϕ_{Λ_D}, corresponding to the lowest encountered $\tilde{C}_D$.

Subsequence Merging, Redecomposition, and Reoptimization. Moving down the decomposition forest, hierarchical ensemble decomposition makes the assumption that base pairs sandwiching parental split-points form with probability approaching unity. Conditional child ensembles enforce these sandwiching base pairs at all levels in the decomposition forest in accordance with the decomposition assumption. As subsequences are merged moving up the decomposition forest, the accuracy of the decomposition assumption is checked. If the assumption is correct, the child-estimated defect will accurately predict the parent-estimated defect. If the assumption is incorrect, the child-estimated defect will not accurately predict the parent-estimated defect since the conditional child ensembles neglect the contributions of structures that lack the sandwiching base pairs. During subsequence merging, if the decomposition assumption is discovered to be incorrect, hierarchical ensemble redecomposition is performed based on the newly available parental base-pairing information. The details of subsequence merging, redecomposition, and reoptimization are as follows.

After leaf reoptimization terminates, parent nodes at depth $d = D - 1$ merge their left and right child sequences to create the candidate sequence $\hat{\phi}_{\Lambda_d}$. The parental test tube ensemble defect estimate, $\tilde{C}_d$, is calculated and the candidate sequence, $\hat{\phi}_{\Lambda_d}$, is accepted if it decreases $\tilde{C}_d$ and rejected otherwise. If the *parental stop condition*

$$\tilde{C}_d \leq \max(C_d^{stop}, \tilde{C}_{d+1}/f_{stringent}) \tag{13}$$

is satisfied, merging continues up to the next level in the forest. Otherwise, failure to satisfy the parental stop condition indicates the existence of the *decomposition defect*,

$$\tilde{C}_d - \tilde{C}_{d+1}/f_{stringent} > 0$$

exceeding the basal level permitted by the parameter $f_{stringent}$. The parent node at depth d whose replacement by its children results in the greatest underestimate of the test tube ensemble defect at level d is subjected to (structure- and probability-guided) hierarchical ensemble redecomposition as described below. Additional parents are redecomposed until

$$\tilde{C}_d - \tilde{C}_{d+1}^*/f_{stringent} \leq f_{redecomp}(\tilde{C}_d - \tilde{C}_{d+1}/f_{stringent})$$

where $\tilde{C}_{d+1}$ is the child defect estimate before any redecomposition, $\tilde{C}_{d+1}^*$ is the child defect estimate after redecomposition, and $f_{redecomp} \in (0, 1)$.

After redecomposition, the current sequences at depth d are pushed to level D, the lowest encountered defect estimate is reset for all levels below d, and a new round of leaf mutation and reoptimization is performed. Following leaf reoptimization, merging begins again. Subsequence merging and reoptimization terminate successfully if the parental stop condition (eq 13) is satisfied at depth $d = 1$. The outcome of subsequence merging, re-decomposition, and reoptimization is the sequence, ϕ_{Λ_1}, corresponding to the lowest encountered $\tilde{C}_1$.

Test Tube Evaluation, Refocusing, and Reoptimization. Using test tube ensemble focusing, initial sequence optimization is performed for the on-target complexes in Ψ^{active}, neglecting the off-target complexes in $\Psi^{passive}$. At the termination of initial forest optimization, the estimated test tube ensemble defect is $\tilde{C}_1$, calculated using eq 11. For this estimate, the complex defect contributions, $\tilde{c}_{\Psi^{active}}$, are based on complex concentration estimates, $\tilde{x}_{\Psi^{active}}$, calculated using deflated total strand concentrations (eq 10) to create a built-in defect allowance for the effect of the neglected off-targets in $\Psi^{passive}$. The exact test tube ensemble defect, C, is then evaluated for the first time over the full ensemble, Ψ, using eq 3. For this exact calculation, the complex defect contributions, c_Ψ, are based on complex concentrations, x_Ψ, calculated using the full strand concentrations (eq 9).

If the test tube ensemble defect satisfies the *termination stop condition*,

$$C \leq \max(C_{stop}, \tilde{C}_1) \tag{14}$$

sequence design terminates successfully. Otherwise, failure to satisfy the termination stop condition indicates the existence of the *focusing defect*:

$$C - \tilde{C}_1 > 0.$$

The test tube ensemble is refocused by transferring the highest-concentration off-target in $\Psi^{passive}$ to Ψ^{active}. Additional off-targets are transferred from $\Psi^{passive}$ to Ψ^{active} until

$$C - \tilde{C}_1^* \leq f_{refocus}(C - \tilde{C}_1)$$

where $\tilde{C}_1$ is the forest-estimated defect before any refocusing, $\tilde{C}_1^*$ is the forest-estimated defect after refocusing (calculated using deflated strand concentrations (eq 10) if $\Psi^{passive} \neq \varnothing$), and $f_{refocus} \in (0, 1)$.

The new off-target structures in Ψ^{active} are then decomposed using probability-guided hierarchical ensemble decomposition as described below, the decomposition forest is augmented with new nodes at all depths, and forest reoptimization commences starting from the final sequences from the previous round of forest optimization. During forest reoptimization, the algorithm

actively attempts to destabilize the off-targets that were added to Ψ^{active}. This process of ensemble refocusing and forest reoptimization is repeated until the termination stop condition (eq 14) is satisfied, which is guaranteed to occur in the event that all off-targets are eventually added to Ψ^{active}. At the conclusion of sequence design, the algorithm returns the sequence set, ϕ_Ψ, that yielded the lowest encountered test tube ensemble defect, C.

Hierarchical Ensemble Decomposition Using Multiple Exclusive Split-Points. Prior to sequence design, in the absence of base-pairing probability information, hierarchical ensemble decomposition was performed for each complex, $j \in \Psi^{\text{active}}$, based on the user-specified on-target structure, s_j. For a parent node, k, with structural ensemble, Γ_k, a single split-point, F, was positioned within a duplex in target structure, s_j, so as to minimize the cost of evaluating both children, yielding left and right child nodes k_l and k_r with conditional ensembles $\tilde{\Gamma}_{k_l}$ and $\tilde{\Gamma}_{k_r}$. These child ensembles enable reconstruction of the conditional parent ensemble, $\tilde{\Gamma}_k$, containing all structures in Γ_k that contain the two base pairs that sandwich F. Following leaf optimization, when left and right child sequences are merged to form a parent sequence, if decomposition defects are observed, $\tilde{\Gamma}_k$ excludes structures that are important to the equilibrium physical properties of Γ_k, implying that the base pairs sandwiching F do not form with probability approaching unity and hence that the conditional physical quantities calculated for the children are not sufficient to predict the physical quantities for the parent. This situation is remedied by redecomposing the parent, taking into consideration the newly available parental base-pairing probabilities.

Two candidate split-points, F_i and F_j, are *exclusive* if they cross when depicted in a polymer graph (denoted $F_i \otimes F_j$; see Figure 2b). The parent ensembles, $\tilde{\Gamma}_{k_i}$ and $\tilde{\Gamma}_{k_j}$, reconstructed from the child ensembles implied by exclusive splits points, F_i and F_j, have no structures in common ($\tilde{\Gamma}_{k_i} \cap \tilde{\Gamma}_{k_j} = \varnothing$). A set of mutually exclusive split-points, $\{F\}$, can be used to non-redundantly decompose the parent ensemble so that the sum of the probabilities of the sandwiching base-pair stacks approaches unity from below. During subsequence merging, redecomposition of parent nodes derived from on-target complexes is performed using structure- and probability-guided decomposition using multiple exclusive split-points. During off-target destabilization, decomposition of parent nodes derived from off-target complexes (for which no target structures exist), is performed using probability-guided decomposition using multiple exclusive split-points. In either case, selection of the optimal set of exclusive split-points is determined using a branch and bound algorithm to minimize the cost of evaluating the child nodes (see Supporting Information section S1.4). Because exclusive split-points lead to exclusive structural ensembles, it is straightforward to generalize the expressions used to reconstruct parental physical properties from child physical properties, as detailed below.

Structure-Guided Decomposition Using a Single Split-Point. For comparison with the formulations that follow, here, we recast the previously described structure-guided hierarchical ensemble decomposition using modified notation. Let F denote a split-point and let $F^{\pm}$ denote the union of the sets of H_{split} base pairs sandwiching F on either side. For a node k descendant from on-target complex $j \in \Psi^{\text{active}}$ with user-specified target structure s_j, the nodal target structure matrix, S_k, is defined using the corresponding entries from the root target structure matrix S_j. The set of valid split-points may be denoted

$$B(S_k) \equiv \left\{ F: \begin{array}{c} \min\limits_{a\cdot b \in F^{\pm}} S_k^{a,b} = 1 \\ \min(|\phi_{k_l}|, |\phi_{k_r}|) \geq N_{\text{split}} \end{array} \right\} \tag{15}$$

and the optimal split-point selected for decomposition,

$$F^* \equiv \min_{F \in B(S_k)} (|\phi_{k_l}|^3 + |\phi_{k_r}|^3)$$

minimizes the cost of evaluating the two child nodes implied by F.

Probability-Guided Decomposition Using Multiple Exclusive Split-Points. The set of sets of valid exclusive split-points may be denoted

$$\bar{B}(P_k) \equiv \left\{ \{F\}: \begin{array}{c} f_{\text{split}} \leq \sum\limits_{F_i \in \{F\}} \min\limits_{a\cdot b \in F_i^{\pm}} P_k^{a,b} \\ \min\limits_{F_i \in \{F\}} (|\phi_{k_{l_i}}|, |\phi_{k_{r_i}}|) \geq N_{\text{split}} \\ F_i \otimes F_j \ \forall F_i \neq F_j \in \{F\} \end{array} \right\} \tag{16}$$

for a user-specified value of $f_{\text{split}} \in (0,1)$ and the optimal set of exclusive split-points selected for decomposition,

$$\{F\}^* \equiv \min_{\{F\} \in \bar{B}(P_k)} \sum_{F_i \in \{F\}} (|\phi_{k_{l_i}}|^3 + |\phi_{k_{r_i}}|^3)$$

minimizes the cost of evaluating the $2|\{F\}|$ child nodes implied by $\{F\}$.

Structure- and Probability-Guided Decomposition using Multiple Exclusive Split-Points. The set of sets of valid split-points may be denoted

$$\hat{B}(S_k, P_k) \equiv \left\{ \{F\}: \begin{array}{c} \{F\} = G_i \cup \{G\}_j,\ G_i \in B(S_k),\ \{G\}_j \in \bar{B}(P_k) \\ F_i \otimes F_j \ \forall F_i \neq F_j \in \{F\} \end{array} \right\}$$

and the optimal set of exclusive split-points selected for decomposition

$$\{F\}^* \equiv \min_{\{F\} \in \hat{B}(S_k, P_k)} \sum_{F_i \in \{F\}} (|\phi_{k_{l_i}}|^3 + |\phi_{k_{r_i}}|^3)$$

minimizes the cost of evaluating the $2|\{F\}|$ child nodes implied by $\{F\}$. The structure-guided component of this approach ensures that the redecomposition is compatible with the user-specified target structure, while the probability-guided component of this approach ensures that the physical properties of the parent can be accurately estimated using the children. For any children resulting from split-points that are target-structure-incompatible, subsequent decomposition is performed via probability-guided decomposition.

Test Tube Ensemble Defect Estimation using Multiple Exclusive Decompositions. Here, we generalize the formulation of test tube ensemble defect estimation at depth $d \in \{2,...,D\}$ to account for the possibility of multiple exclusive split-points within any parent in the decomposition forest. Consider parent k decomposed using the set of exclusive split-points, $\{F\}$.

Following eq 7, for each split-point $F_i \in \{F\}$, the corresponding exclusive contribution to the parental partition function estimate is

dx.doi.org/10.1021/sb5002196 | ACS Synth. Biol. 2015, 4, 1086−1100

$$\tilde{Q}_{k_i} = \tilde{Q}_{k_{l_i}} \tilde{Q}_{k_{r_i}} \exp(-\Delta G_{F_i}^{\text{interior}}/k_B T)$$

yielding the partition function estimate for parent k:

$$\tilde{Q}_k = \sum_{F_i \in \{F\}} \tilde{Q}_{k_i}$$

Recursive merging is performed until the complex partition function estimate, $\tilde{Q}_j$, is obtained.

Likewise, for each split-point $F_i \in \{F\}$, the entries in the exclusive parental pair probability matrix estimate, $\tilde{P}_{k_i}$, are taken from the corresponding entries in the child pair probability matrix estimates, $\tilde{P}_{k_{l_i}}$ and $\tilde{P}_{k_{r_i}}$. Boltzmann weighting of exclusive parental estimates then yields the pair probability matrix estimate for parent k:

$$\tilde{P}_k = \sum_{F_i \in \{F\}} \tilde{P}_{k_i} \frac{\tilde{Q}_{k_i}}{\tilde{Q}_k}$$

Recursive merging is performed until the complex pair probability matrix estimate, $\tilde{P}_j$, is obtained.

Calculation of the complex concentration estimates, $\tilde{x}_{\Psi^{\text{active}}}$, the complex ensemble defect estimates, $\tilde{n}_{\Psi^{\text{on}}}$, and the test tube ensemble defect estimate, $\tilde{C}_d$, then proceed as before.

METHODS

Implementation. The test tube design algorithm is coded in the C programming language. The algorithm is available for noncommercial research purposes as part of the NUPACK web application and code base (www.nupack.org).[12]

Target Test Tubes. Algorithm performance is demonstrated using test sets of target test tubes. For the *engineered test set*, each on-target structure was randomly generated with stem and loop sizes randomly selected from a distribution of sizes representative of the nucleic acid engineering literature (see Supporting Information section S3). For the *random test set*, each on-target structure was generated by calculating the minimum free energy structure of a different random RNA sequence at 37 °C. Within each target test tube, there are two on-target dimers (each with a target concentration of 1 μM) and a total of 106 off-target monomers, dimers, trimers, and tetramers (each with vanishing target concentration), representing all complexes of up to $L_{\max} = 4$ strands (excluding the two on-target dimers). For each test set, 50 target test tubes were generated for each on-target dimer size, $|\phi| \in \{50, 100, 200, 400\}$ nt, with all strands the same length in each target test tube. The structural properties of the on-target structures in the engineered and random test sets are summarized in Supporting Information Figure S18. Typically, the random test set contains on-target structures with a lower fraction of paired nucleotides, more stems, and shorter stems (as short as one base pair). For the design studies that follow, new target test tubes were generated from scratch. The design algorithm was not tested on these target test tubes prior to generating the depicted results.

Sequence Design Trials. For all studies, five independent design trials were performed for each target test tube. Design trials were run on a cluster of 2.53 GHz Intel E5540 Xeon dual-processor/quad-core nodes with 24 GB of memory per node. Each trial was run on a single computational core using the default algorithm parameters of Table 1 unless otherwise noted. Design quality is plotted[62] as the *normalized test tube ensemble defect*, C/y_{nt}. Data are typically presented as cumulative histograms over design trials. Our primary test scenario is RNA sequence design for the engineered test set at 37 °C with $f_{\text{stop}} = 0.01$ (i.e., no more than 1% of the nucleotides in the test tube incorrectly paired at equilibrium).

Table 1. Default Parameters for RNA Design[a]

parameter	value
f_{stop}	0.01
f_{passive}	0.01
H_{split}	2
N_{split}	12
f_{split}	0.99
$f_{\text{stringent}}$	0.99
ΔG^{clamp}	−25 kcal/mol
M_{bad}	300
M_{reseed}	50
M_{reopt}	3
f_{redecomp}	0.03
f_{refocus}	0.03

[a]For DNA design, $H_{\text{split}} = 3$.

RESULTS AND DISCUSSION

Algorithm Performance for Test Tube Design. Figure 4 demonstrates the performance of the test tube design algorithm on the engineered and random test sets. For each target test tube, the algorithm designs for two on-target dimers (each with a target secondary structure and target concentration) and against a total of 106 off-target monomers, dimers, trimers, and tetramers (each with vanishing target concentration). Most designs trials surpass the desired design quality (normalized test tube ensemble defect ≤0.01; panel a). Typical design cost ranges from seconds for test tubes with 50-nt on-targets to hours for test tubes with 400-nt on-targets (panel b). Starting from random initial sequences, the desired design quality can be achieved with a broad range of GC contents (panel c). The typical cost of test tube design relative to the cost of a single evaluation of the test tube ensemble defect is only a factor of 3 for the engineered test set and a factor of 10 for the random test set.

Importance of Designing Against Off-Targets. Is it important to include off-target complexes in the test tube ensemble so that the design algorithm can actively destabilize them? To examine this question in the context of the engineered test set, Figure 5 compares design quality for sequences designed in a test tube ensemble containing either: no off-targets (equivalent to complex design), all off-targets up to dimers, or all off-targets up to tetramers. The quality of the resulting design is evaluated using a reference test tube ensemble including all off-targets up to pentamers. Only the sequences designed against all off-targets up to tetrameters are consistently of high quality; designing against no off-targets or against all off-targets up to dimers typically leads to very poor sequence designs.

We note that, in the absence of off-target destabilization, the strands in an on-target complex will often form a dimerized off-target complex (containing two copies of each strand) at significant concentrations. We recommend actively designing against these dimerized off-targets, which is achieved for the engineered test set (containing dimer on-targets) by designing against all off-targets up to tetramers.

Contributions of Algorithmic Ingredients. To avoid the expense of evaluating candidate mutations on all off-target complexes throughout the design process, test tube ensemble

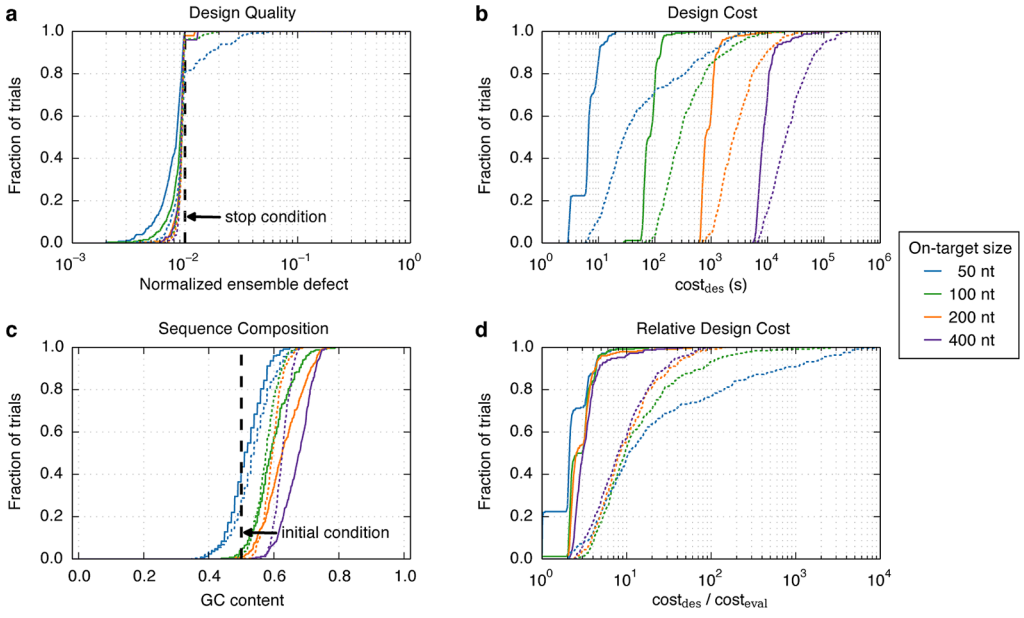


Figure 4. Algorithm performance for test tube design. (a) Design quality. The stop condition is depicted as a dashed black line. (b) Design cost. (c) Sequence composition. The initial GC content is depicted as a dashed black line. (d) Cost of sequence design relative to a single evaluation of the objective function. RNA design at 37 °C for the engineered test set (solid lines) and random test set (dotted lines).

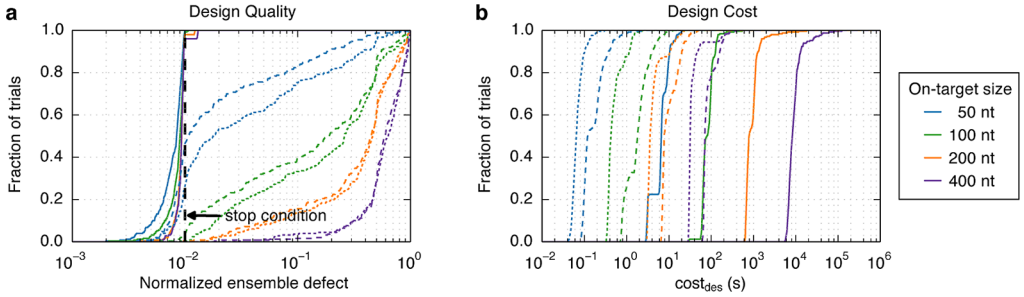


Figure 5. Importance of designing against off-targets. Comparison of test tube design performed using an ensemble containing all off-targets up to size L_{max} = 0 (dotted line; $|\Psi^{off}|$ = 0, equivalent to complex design), L_{max} = 2 (dashed line; $|\Psi^{off}|$ = 12), or L_{max} = 4 (solid line; $|\Psi^{off}|$ = 106). (a) Design quality evaluated by calculating the test tube ensemble defect for a reference ensemble containing all off-targets up to size L_{max} = 5 ($|\Psi^{off}|$ = 314). The stop condition is depicted as a dashed black line. (b) Design cost. RNA design at 37 °C for the engineered test set.

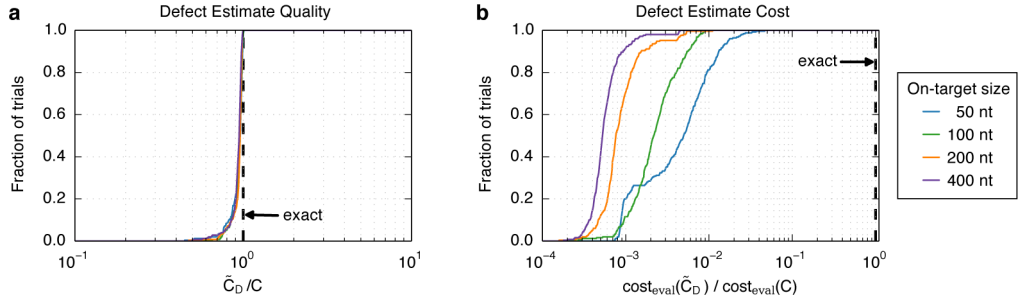


Figure 6. Accuracy and cost of test tube ensemble defect estimation. (a) Accuracy of leaf-based estimate relative to cost of exact defect. (b) Cost of leaf-based estimate relative to exact defect cost. The exact test tube ensemble defect, C, is calculated using all on- and off-target complexes $j \in \Psi$. The test tube ensemble defect estimate, $\tilde{C}_D$, is calculated using the leaf nodes, $k \in \Lambda_D$, of the final decomposition forest obtained by hierarchically decomposing the structural ensembles of the on- and off-target complexes $j \in \Psi^{active}$. RNA design at 37 °C for the engineered test set. For each design trial, comparisons are made using the final designed sequences, ϕ_Ψ.

focusing partitions the complexes in Ψ into the sets Ψ^{active} and $\Psi^{passive}$. To efficiently accept or reject each candidate mutation, the test tube ensemble defect is estimated at the leaf level of the decomposition forest obtained via hierarchical ensemble decomposition of the complexes in Ψ^{active}. Figure 6 demonstrates that, for the engineered test set, the estimated defect typically closely approximates the exact defect (panel a) but at a cost that is lower by 2−3 orders of magnitude (panel b). Figure 7 demonstrates that the cost savings resulting from hierarchical ensemble decomposition become substantial as the size of the complexes in Ψ^{active} increases (due to the increasing depth of the decomposition forest) and that the cost savings

dx.doi.org/10.1021/sb5002196 | ACS Synth. Biol. 2015, 4, 1086−1100

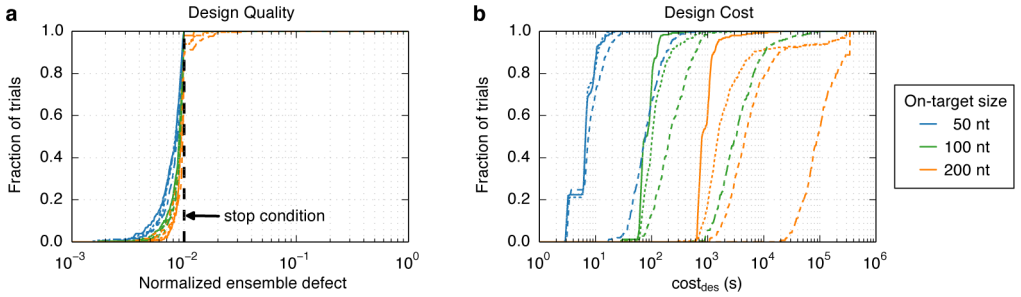


Figure 7. Efficiency implications of test tube ensemble focusing and hierarchical ensemble decomposition. (a) Design quality. The stop condition is depicted as a dashed black line. (b) Design cost. Comparison of test tube design performed with the full algorithm (including test tube ensemble focusing and hierarchical ensemble decomposition permitting multiple exclusive split-points per parent; solid lines), test tube ensemble focusing and hierarchal ensemble decomposition permitting only a single split-point per parent (dotted lines), test tube ensemble focusing but no hierarchical ensemble decomposition (dashed lines), or no test tube ensemble focusing and no hierarchical ensemble decomposition (uneven dashed lines). RNA design at 37 °C for the subset of the engineered test set with 50-nt, 100-nt, and 200-nt on-targets.

resulting from test tube ensemble focusing are substantial independent of complex size (due to the large number of off-targets in Ψ^{passive}).

Robustness of Predictions to Model Perturbations. Algorithms for the analysis and design of equilibrium nucleic acid secondary structure depend on empirical free energy models.[30−34] It is inevitable that the parameter sets in these models will continue to be refined, so it is important that assessments of design quality are robust to parameter perturbations. Figure 8 demonstrates that for the subset of

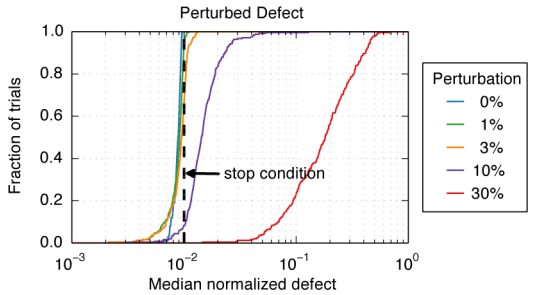


Figure 8. Robustness of design quality predictions to perturbations in model parameters. For each design trial, the median test tube ensemble defect was calculated over 100 perturbed physical models (each parameter perturbed by Gaussian noise with a standard deviation of 0, 1, 3, 10, or 30% of the parameter modulus). RNA design at 37 °C for the subset of the engineered test set with 100-nt on-targets. The stop condition is depicted as a dashed black line.

the engineered test set with 100-nt on-targets, the predicted quality of most sequence designs is typically robust to 3% parameter perturbations (with test tube ensemble defect often less than the stop condition), and even to 10% parameter perturbations (with test tube ensemble defect typically within a factor of 2 of the stop condition), but not to 30% parameter perturbations (with test tube ensemble defect rarely within a factor of 10 of the stop condition).

Test Tube Design with Competing On-Target Complexes. In the engineered test set, each of four strands appears in exactly one of two on-target dimers so there is no disadvantage to stabilizing these dimers to the maximum extent possible since the target concentration for all off-target complexes is zero. However, if there are multiple on-target complexes competing for the same strands, then the algorithm must balance the relative stability of these competing on-targets. To examine this challenge, we consider target test tubes in which a strand is intended to form both a monomer hairpin and a dimer duplex (Figure 9a), varying the target concentration of the monomer while keeping the total strand concentration fixed. Figure 9b demonstrates that typical design quality varies greatly depending on the desired relative stability of the monomer and dimer on-targets. For example, the algorithm typically succeeds in satisfying the stop condition for low monomer target concentrations but not for high monomer target concentrations. These designs were performed using Watson−Crick complementarity constraints. If wobble pairs are permitted, typical design performance significantly improves (Figure 9c), reflecting the additional flexibility provided to the algorithm.

Because of the competition between on-target complexes, it is interesting to revisit the question of robustness to model perturbations. The perturbation studies of Figure 9d demonstrate that the predicted design quality is typically robust to model perturbations for test tubes where one on-target dominates the other but becomes more sensitive to model perturbations for test tubes where both on-targets are in competition at nonsaturated target concentrations. Hence, for applications where multiple competing on-targets are intended to form at nonsaturated concentrations, it is more likely that the relative stabilities of the on-targets will need to be fine-tuned based on experimental measurements to account for imperfections in the physical model. Fortunately, many applications seek to saturate all on-targets at maximum concentration, reducing the sensitivity of computational predictions to perturbations in the model parameters.

Test Tube Design with Large Numbers of On- and Off-Targets. Figure 10 demonstrates the performance of the algorithm for target test tubes containing large number of on- and off-target complexes. Typical design trials surpass the desired design quality (normalized test tube ensemble defect $\leq$0.01; panel a) and the typical design cost is less than 4 times the cost of a single evaluation of the test tube ensemble defect (panel d), ranging from 10 s for a test tube containing 1 on-target and 14 off-targets to 8 h for a test tube containing 8 on-targets and 17976 off-targets (panel b).

CONCLUSION

Test tube design provides a powerful framework for engineering nucleic acid base pairing. The desired equilibrium base-pairing properties for the test tube are specified as an arbitrary number of on-target complexes, each with a target secondary structure and target concentration, and an arbitrary number of

dx.doi.org/10.1021/sb5002196 | ACS Synth. Biol. 2015, 4, 1086−1100

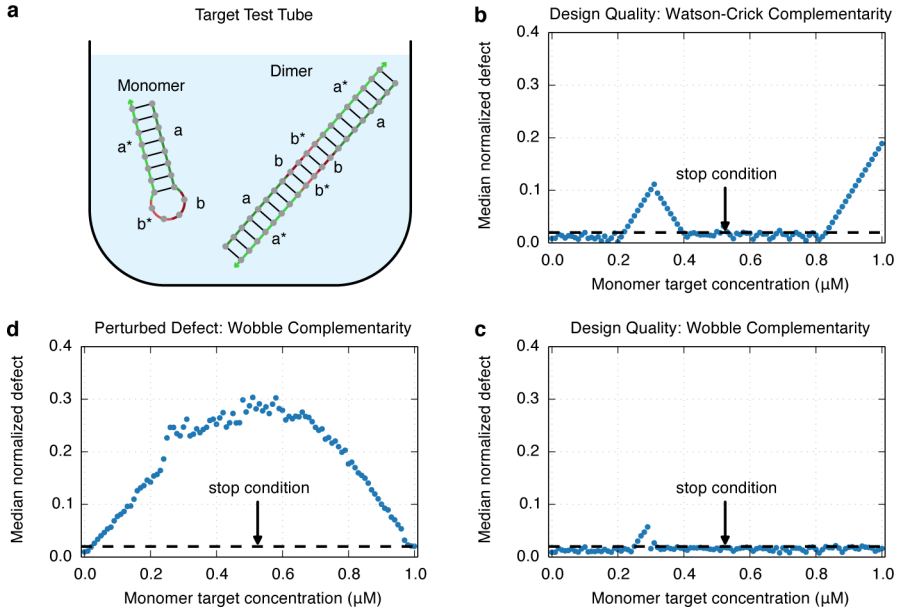


Figure 9. Test tube design with competing on-target complexes. (a) A range of target test tubes were defined with the monomer target concentration ranging from 0 to 1 μM in 0.01 μM increments and the total strand concentration held fixed at 1 μM. (b) Median design quality with Watson−Crick complementarity constraints. (c) Median design quality with wobble complementarity constraints. (d) Robustness of design quality predictions to perturbations in model parameters for sequence designs with wobble complementarity constraints. For each design trial, the median test tube ensemble defect was calculated over 100 perturbed physical models (each parameter perturbed by Gaussian noise with a standard deviation of 10% of the parameter modulus). RNA sequence design at 37 °C with L_{max} = 2 (i.e., no off-targets). The test tube stop condition is depicted as a dashed black line (f_{stop} = 0.02).

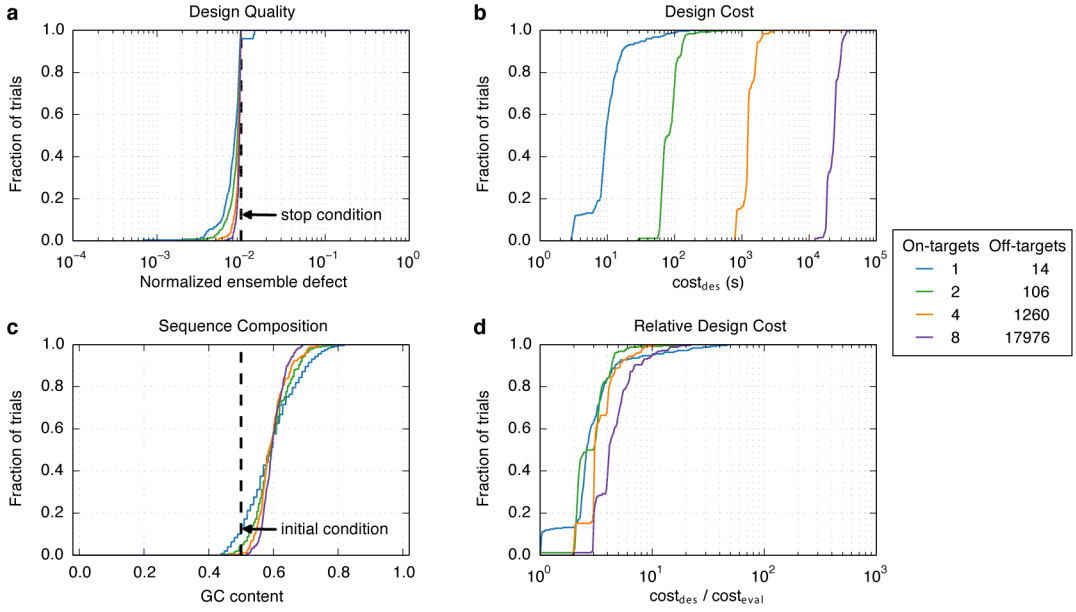


Figure 10. Test tube design with large numbers of on- and off-target complexes. Target test tubes contain $|\Psi^{on}| \in \{1,2,4,8\}$ on-target dimers and all off-target complexes up to size L_{max} = 4 (corresponding to target test tubes with $|\Psi^{off}| \in \{14,106,1260,17976\}$, respectively). (a) Design quality. The stop condition is depicted as a dashed black line. (b) Design cost. (c) Sequence composition. The initial GC content is depicted as a dashed black line. (d) Cost of sequence design relative to a single evaluation of the objective function. RNA sequence design at 37 °C. Fifty target test tubes for each value of $|\Psi^{on}|$; 100-nt on-target dimers randomly selected from the engineered test set.

off-target complexes, each with vanishing target concentration. Given a target test tube, the test tube ensemble defect quantifies the concentration of incorrectly paired nucleotides at equilibrium. Test tube ensemble defect optimization implements a positive design paradigm (stabilize on-targets) and a negative design paradigm (destabilize off-targets) at two levels: (a) designing for the on-target structure and against all off-target structures within the structural ensemble of each on-target complex,[15,22] and (b) designing for the target concentration of each on-target complex and against the formation of all off-target complexes within the ensemble of the test tube. Both paradigms are crucial at both levels in order to achieve high-quality test tube designs with a low test tube ensemble defect.

dx.doi.org/10.1021/sb5002196 | ACS Synth. Biol. 2015, 4, 1086−1100

Three concepts enable efficient test tube design by reducing the cost of evaluating candidate sequences:

- *Test tube ensemble focusing:* Test tube ensemble focusing dramatically reduces the number of complexes that are actively considered within the ensemble of the test tube during sequence optimization. Initially, only on-target complexes are included in the focused ensemble, making the assumption that all off-target complexes will form with negligible concentration at equilibrium. If this assumption proves incorrect, test tube ensemble refocusing is performed to ensure that any off-target complexes observed to form with non-negligible concentration in the full test tube ensemble are included in the focused ensemble for active destabilization during subsequent rounds of sequence optimization.
- *Hierarchical ensemble decomposition:* Hierarchical ensemble decomposition dramatically reduces the number of structures that are actively considered within the ensemble of each complex during sequence optimization. The structural ensemble of each complex in the focused test tube ensemble is decomposed into a tree of conditional subensembles, yielding a forest of decomposition trees. Initially, decomposition of each parent ensemble is based on the assumption that many off-target structures that are incompatible with the on-target structure for the complex will form with negligible probability and may be neglected in the conditional child ensembles that are used to efficiently estimate physical properties of the parent. If this assumption proves incorrect, hierarchical ensemble redecomposition is performed to ensure that any off-target structures observed to form with non-negligible probability in the parent ensemble are included in the conditional child ensembles for active destabilization during subsequent rounds of sequence optimization.
- *Calculation of conditional physical properties:* By calculating conditional partition functions and pair probabilities over the conditional structural ensembles at the leaves of the decomposition forest, the equilibrium base-pairing properties of a test tube of interacting strands can be accurately and efficiently estimated. This estimation capability applies to the test tube ensemble defect that provides a physically meaningful objective function for test tube design, as well as to other underlying physical properties (complex partition functions, complex pair probabilities, and complex concentrations) that may be of interest in other contexts.

Used in combination, test tube ensemble focusing, hierarchical ensemble decomposition, and calculation of conditional physical properties enable efficient estimation and optimization of the test tube ensemble defect for target test tubes representative of design challenges in the molecular programming and synthetic biology communities, typically achieving a normalized test tube ensemble defect $\leq$1% at a design cost within an order of magnitude of the cost of test tube analysis.

■ ASSOCIATED CONTENT

Supporting Information

Additional algorithm details: pseudocode, calculation of conditional partition functions and base-pairing probabilities, distinguishability issues, selection and use of multiple exclusive split-points. Additional design studies: sensitivity of performance to algorithm parameters, complex design, sequence initialization and reseeding, RNA vs DNA design. Structural features of the engineered and random test sets. On-target secondary structures for all test sets. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*Email: niles@caltech.edu.

Author Contributions

B.R.W. and N.A.P. developed the algorithm, designed the computational experiments, and wrote the paper. B.R.W. wrote the software and performed the computational experiments.

Notes

The authors declare the following competing financial interest(s): Pending patent applications.

■ ACKNOWLEDGMENTS

The authors thank J. S. Bois, J. N. Zadeh, and N. J. Porubsky for helpful discussions and M. Kirk for assistance with bibliography data entry. This work was funded by the National Science Foundation via the Molecular Programming Project (NSF-CCF-0832824 and NSF-CCF-1317694), by the Gordon and Betty Moore Foundation (GBMF2809), by the John Simon Guggenheim Memorial Foundation, and by the Beckman Institute at Caltech.

■ REFERENCES

(1) Pinheiro, A. V., Han, D. R., Shih, W. M., and Yan, H. (2011) Challenges and opportunities for structural DNA nanotechnology. *Nat. Nanotechnol. 6*, 763−772.

(2) Zhang, D. Y., and Seelig, G. (2011) Dynamic DNA nanotechnology using strand-displacement reactions. *Nat. Chem. 3*, 103−113.

(3) Zuker, M., and Stiegler, P. (1981) Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res. 9*, 133−147.

(4) McCaskill, J. S. (1990) The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers 29*, 1105−1119.

(5) 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. *Monatsh. Chem. 125*, 167−188.

(6) Lyngso, R. B., Zuker, M., and Pedersen, C. N. S. (1999) Fast evaluation of internal loops in RNA secondary structure prediction. *Bioinformatics 15*, 440−445.

(7) Dirks, R. M., and Pierce, N. A. (2004) An algorithm for computing nucleic acid base-pairing probabilities including pseudoknots. *J. Comput. Chem. 25*, 1295−1304.

(8) Dimitrov, R. A., and Zuker, M. (2004) Prediction of hybridization and melting for double-stranded nucleic acids. *Biophys. J. 87*, 215−226.

(9) Andronescu, M., Zhang, Z. C., and Condon, A. (2005) Secondary structure prediction of interacting RNA molecules. *J. Mol. Biol. 345*, 987−1001.

(10) Bernhart, S. H., Tafer, H., Muckstein, U., Flamm, C., Stadler, P. F., and Hofacker, I. L. (2006) Partition function and base pairing probabilities of RNA heterodimers. *Algorithm Mol. Biol. 1*, 3 DOI: 10.1186/1748-7188-1-3.

(11) Dirks, R. M., Bois, J. S., Schaeffer, J. M., Winfree, E., and Pierce, N. A. (2007) Thermodynamic analysis of interacting nucleic acid strands. *SIAM Rev. 49*, 65−88.

(12) Zadeh, J. N., Steenberg, C. D., Bois, J. S., Wolfe, B. R., Pierce, M. B., Khan, A. R., Dirks, R. M., and Pierce, N. A. (2011) NUPACK:

Analysis and design of nucleic acid systems. *J. Comput. Chem. 32*, 170–173.
(13) Flamm, C., Hofacker, I. L., Maurer-Stroh, S., Stadler, P. F., and Zehl, M. (2001) Design of multistable RNA molecules. *RNA 7*, 254–265.
(14) 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.
(15) Dirks, R. M., Lin, M., Winfree, E., and Pierce, N. A. (2004) Paradigms for computational nucleic acid design. *Nucleic Acids Res. 32*, 1392–1403.
(16) Andronescu, M., Fejes, A. P., Hutter, F., Hoos, H. H., and Condon, A. (2004) A new algorithm for RNA secondary structure design. *J. Mol. Biol. 336*, 607–624.
(17) Busch, A., and Backofen, R. (2006) INFO-RNA—A fast approach to inverse RNA folding. *Bioinformatics 22*, 1823–1831.
(18) Aguirre-Hernandez, R., Hoos, H. H., and Condon, A. (2007) Computational RNA secondary structure design: Empirical complexity and improved methods. *BMC Bioinformatics 8*, 34 DOI: 10.1186/1471-2105-8-34.
(19) Burghardt, B., and Hartmann, A. K. (2007) RNA secondary structure design. *Phys. Rev. E 75*, 021920.
(20) Gao, J. Z. M., Li, L. Y. M., and Reidys, C. M. (2010) Inverse folding of RNA pseudoknot structures. *Algorithm Mol. Biol. 5*, 27 DOI: 10.1186/1748-7188-5-27.
(21) Shu, W. J., Liu, M., Chen, H. B., Bo, X. C., and Wang, S. Q. (2010) ARDesigner: A web-based system for allosteric RNA design. *J. Biotechnol. 150*, 466–473.
(22) Zadeh, J. N., Wolfe, B. R., and Pierce, N. A. (2011) Nucleic acid sequence design via efficient ensemble defect optimization. *J. Comput. Chem. 32*, 439–452.
(23) Avihoo, A., Churkin, A., and Barash, D. (2011) RNAexinv: An extended inverse RNA folding from shape and physical attributes to sequences. *BMC Bioinformatics 12*, 319 DOI: 10.1186/1471-2105-12-319.
(24) Ramlan, E. I., and Zauner, K. P. (2011) Design of interacting multi-stable nucleic acids for molecular information processing. *Biosystems 105*, 14–24.
(25) Taneda, A. (2011) MODENA: A multi-objective RNA inverse folding. *Adv. Appl. Bioinforma. Chem. 4*, 1–12.
(26) Levin, A., Lis, M., Ponty, Y., O'Donnell, C. W., Devadas, S., Berger, B., and Waldispuhl, J. (2012) A global sampling approach to designing and reengineering RNA secondary structures. *Nucleic Acids Res. 40*, 10041–10052.
(27) Matthies, M. C., Bienert, S., and Torda, A. E. (2012) Dynamics in sequence space for RNA secondary structure design. *J. Chem. Theory Comput. 8*, 3663–3670.
(28) Taneda, A. (2012) Multi-objective genetic algorithm for pseudoknotted RNA sequence design. *Front. Genet. 3*, 36.
(29) Lyngso, R. B., Anderson, J. W. J., Sizikova, E., Badugu, A., Hyland, T., and Hein, J. (2012) Frnakenstein: Multiple target inverse RNA folding. *BMC Bioinformatics 13*, 260 DOI: 10.1186/1471-2105-13-260.
(30) Serra, M. J., and Turner, D. H. (1995) Predicting thermodynamic properties of RNA. *Methods Enzymol. 259*, 242–261.
(31) 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.
(32) SantaLucia, J., Jr. (1998) A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc. Natl. Acad. Sci. U.S.A. 95*, 1460–1465.
(33) SantaLucia, J., Jr., and Hicks, D. (2004) The thermodynamics of DNA structural motifs. *Annu. Rev. Biophys. Biomol. Struct. 33*, 415–440.
(34) Koehler, R. T., and Peyret, N. (2005) Thermodynamic properties of DNA sequences: Characteristic values for the human genome. *Bioinformatics 21*, 3333–3339.
(35) Genot, A. J., Zhang, D. Y., Bath, J., and Turberfield, A. J. (2011) Remote toehold: A mechanism for flexible control of DNA hybridization kinetics. *J. Am. Chem. Soc. 133*, 2177–2182.
(36) Genot, A. J., Bath, J., and Turberfield, A. J. (2011) Reversible logic circuits made of DNA. *J. Am. Chem. Soc. 133*, 20080–20083.
(37) Delebecque, C. J., Silver, P. A., and Lindner, A. B. (2012) Designing and using RNA scaffolds to assemble proteins *in vivo*. *Nat. Protoc. 7*, 1797–1807.
(38) Greene, D. G., Keum, J. W., and Bermudez, H. (2012) The role of defects on the assembly and stability of DNA nanostructures. *Small 8*, 1320–1325.
(39) Padirac, A., Fujii, T., and Rondelez, Y. (2012) Quencher-free multiplexed monitoring of DNA reaction circuits. *Nucleic Acids Res. 40*, e118.
(40) Tang, H., Deschner, R., Allen, P., Cho, Y. J., Sermas, P., Maurer, A., Ellington, A. D., and Willson, C. G. (2012) Analysis of DNA-guided self-assembly of microspheres using imaging flow cytometry. *J. Am. Chem. Soc. 134*, 15245–15248.
(41) Zhang, X. J., and Yadavalli, V. K. (2012) Functional self-assembled DNA nanostructures for molecular recognition. *Nanoscale 4*, 2439–2446.
(42) Goodman, D. B., Church, G. M., and Kosuri, S. (2013) Causes and effects of N-terminal codon bias in bacterial genes. *Science 342*, 475–479.
(43) Xu, X. W., and Yang, X. R. (2014) Reversion of DNA strand displacement using functional nucleic acids as toeholds. *Chem. Commun. 50*, 805–807.
(44) Dirks, R. M., and Pierce, N. A. (2004) Triggered amplification by hybridization chain reaction. *Proc. Natl. Acad. Sci. U.S.A. 101*, 15275–15278.
(45) Patzel, V., Rutz, S., Dietrich, I., Köberle, C., Sheffold, A., and Kaufmann, S. H. E. (2005) Design of siRNAs producing unstructured guide-RNAs results in improved RNA interference efficiency. *Nat. Biotechnol. 23*, 1440–1444.
(46) Penchovsky, R., and Breaker, R. R. (2005) Computational design and experimental validation of oligonucleotide-sensing allosteric ribozymes. *Nat. Biotechnol. 23*, 1424–1433.
(47) Venkataraman, S., Dirks, R. M., Rothemund, P. W. K., Winfree, E., and Pierce, N. A. (2007) An autonomous polymerization motor powered by DNA hybridization. *Nat. Nanotechnol. 2*, 490–494.
(48) Yin, P., Choi, H. M. T., Calvert, C. R., and Pierce, N. A. (2008) Programming biomolecular self-assembly pathways. *Nature 451*, 318–322.
(49) Salis, H. M., Mirsky, E. A., and Voigt, C. A. (2009) Automated design of synthetic ribosome binding sites to control protein expression. *Nat. Biotechnol. 27*, 946–950.
(50) Li, B. L., Ellington, A. D., and Chen, X. (2011) Rational, modular adaptation of enzyme-free DNA circuits to multiple detection methods. *Nucleic Acids Res. 39*, e110.
(51) Dong, J., Cui, X., Deng, Y., and Tang, Z. (2012) Amplified detection of nucleic acid by G-quadruplex based hybridization chain reaction. *Biosens. Bioelectron. 38*, 258–263.
(52) Nishimura, T., Ogura, Y., and Tanida, J. (2012) Fluorescence resonance energy transfer-based molecular logic circuit using a DNA scaffold. *Appl. Phys. Lett. 101*, 233703.
(53) Schade, M., Knoll, A., Vogel, A., Seitz, O., Liebscher, J., Huster, D., Herrmann, A., and Arbuzova, A. (2012) Remote control of lipophilic nucleic acids domain partitioning by DNA hybridization and enzymatic cleavage. *J. Am. Chem. Soc. 134*, 20490–20497.
(54) Vieregg, J. R., Nelson, H. M., Stoltz, B. M., and Pierce, N. A. (2013) Selective nucleic acid capture with shielded covalent probes. *J. Am. Chem. Soc. 135*, 9691–9699.
(55) Hochrein, L. M., Schwarzkopf, M., Shahgholi, M., Yin, P., and Pierce, N. A. (2013) Conditional Dicer substrate formation via shape and sequence transduction with small conditional RNAs. *J. Am. Chem. Soc. 135*, 17322–17330.
(56) Genot, A. J., Bath, J., and Turberfield, A. J. (2013) Combinatorial displacement of DNA strands: Application to matrix

multiplication and weighted sums. *Angew. Chem., Int. Ed. 52*, 1189−1192.
(57) Hamblin, G. D., Hariri, A. A., Carneiro, K. M. M., Lau, K. L., Cosa, G., and Sleiman, H. F. (2013) Simple design for DNA nanotubes from a minimal set of unmodified strands: Rapid, room-temperature assembly, and readily tunable structure. *ACS Nano 7*, 3022−3028.
(58) Santini, C. C., Bath, J., Tyrrell, A. M., and Turberfield, A. J. (2013) A clocked finite state machine built from DNA. *Chem. Commun. 49*, 237−239.
(59) Jiang, Y. S., Bhadra, S., Li, B. L., and Ellington, A. D. (2014) Mismatches improve the performance of strand-displacement nucleic acid circuits. *Angew. Chem., Int. Ed. 53*, 1845−1848.
(60) Geary, C., Rothemund, P. W. K., and Andersen, E. S. (2014) A single-stranded architecture for cotranscriptional folding of RNA nanostructures. *Science 345*, 799−804.
(61) Mathews, D. H. (2004) Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization. *RNA 10*, 1178−1190.
(62) Hunter, J. D. (2007) Matplotlib: A 2D graphics environment. *Comput. Sci. Eng. 9*, 90−95.