DNA Sticky End Design and Assignment for Robust Algorithmic Self-assembly

Constantine G. Evans[1] and Erik Winfree[2]

[1] Physics,
[2] Computer Science,
California Institute of Technology

Abstract. A major challenge in practical DNA tile self-assembly is the minimization of errors. Using the kinetic Tile Assembly Model, a theoretical model of self-assembly, it has been shown that errors can be reduced through abstract tile set design. In this paper, we instead investigate the effects of "sticky end" sequence choices in systems using the kinetic model along with the nearest-neighbor model of DNA interactions. We show that both the sticky end sequences present in a system and their positions in the system can significantly affect error rates, and propose algorithms for sequence design and assignment.

1 Introduction

Self-assembly of DNA tiles is a promising technique for the assembly of complex nanoscale structures. Assembly of tiles can be programmed by designing short complementary single-stranded DNA "sticky ends." While assembly using unique tile types or simple lattices is often studied [26,16], algorithmic growth, where small sets with few tile types can form complex assemblies, is particularly powerful theoretically, and has been studied extensively through the abstract Tile Assembly Model (aTAM) [28,8,17].

A number of different designs for tile structure are used for assembly [26,21,16]. As an example, the DAO-E tile design (Fig. 1(a)) consists of two helices connected by two crossovers, with four 5 nucleotide (nt) sticky ends, one at each end of each helix. Experimentally, conditions are usually used such that tiles will favorably attach by two bonds between sticky-end regions, adding cooperativity to binding. In the abstract Tile Assembly Model, this is modelled by individual tiles attaching to edges of the current assembly when they can make at least two correct bonds to adjacent tiles ($T = 2$), and never detaching once attached.

The Pascal mod 3 (PM3) system shown in Fig. 1(b) is a simple example. The tiles implement addition modulo 3, akin to Pascal's triangle. Tiles attach by their two lower-left ends, and then provide ends for future tiles to attach that sum the logical values of the two "input" ends. Growth proceeds to the upper-right, controlled by a V-shaped seed of tiles that attach by strength-2 bonds and provide edges of logical 1s.

A more sophisticated example, the counter system from Barish et al [3], is shown in Fig. 1(c). In this system, a ribbon of tiles grows from a large seed

D. Soloveichik and B. Yurke (Eds.): DNA 2013, LNCS 8141, pp. 61–75, 2013.

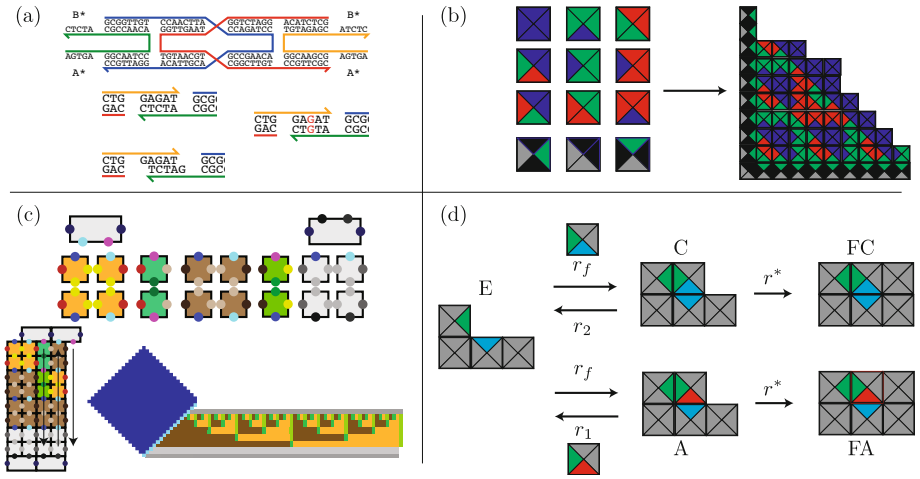

Fig. 1. Tile systems, structures and the kinetic trapping model. (a) shows an example DAO-E tile structure [21], along with examples of complementary and partially mismatched sticky end attachments. (b) shows the Pascal mod 3 tile system along with a potential perfect assembly. Blue, green and red correspond to ends with logical values 0, 1, and 2, respectively, while black indicates double-strength bonds of the V-shaped seed. (c) shows the tiles (top) in the Barish counter system, along with an illustration of zig-zag ribbon growth (left) and an Xgrow simulation of growth from an origami seed (blue), where each pixel represents one tile. Orange and brown tiles indicate tiles with logical values of 1 and 0, respectively, while gray tiles are boundary and nucleation barrier tiles, and incrementing tiles are green. (d) illustrates the states and transition rates in the kinetic trapping model of growth errors.

structure of DNA origami. Rows of tiles grow in a zig-zag fashion, with each new row being started by a double tile that is equivalent to two permanently-attached single tiles. On "downward" rows tiles increment a bit string with two tiles per bit from the previous row, while on "upward" rows, corresponding tiles copy the newly-incremented row. These tiles implement a binary counter starting from whatever bit string was specified on the original origami seed and incrementing every two rows of tiles.

In examining algorithmic growth of experimental systems, the kinetic Tile Assembly Model provides better physical relevance [28]. Tiles are assumed to be in solution at a particular concentration, which is usually assumed to be constant. Tiles attach to empty lattice sites at a rate r_f dependent only on their concentration, and detach at a rate r_b ($b = 1, 2, \ldots$) dependent upon the number of correct "sticky end" attachments they have to the assembly:

$$r_f = \hat{k}e^{-G_{mc}} \qquad r_b = \hat{k}e^{-bG_{se}} \ . \tag{1}$$

Here G_{mc} is a dimensionless free energy analogue related to tile concentration by $[c] = e^{-G_{mc}+\alpha}$, G_{se} is the sign-reversed dimensionless free energy of a single bond, b is the number of correct bonds, and $\hat{k}$ is an adjusted forward rate constant

$\hat{k} \equiv k_f e^{\alpha}$, where k_f is the usual second-order mass action rate constant for tile attachment, typically $k_f = 10^6$ /M/s. This model has been used for numerous theoretical and computational simulation studies of algorithmic tile assembly [29,6,10,8,17], and has fit well with experimental findings both qualitatively and quantitatively [9,11].

As the kinetic model allows any tile to attach regardless of correctness, it is challenging to design tile systems that exhibit algorithmic behavior while keeping erroneous growth low enough to obtain high yields of correct assemblies. Growth errors in the kinetic model are well studied, and often modelled by the kinetic trapping model. The model considers tiles attaching and detaching at a single lattice location, while having a rate for an attached tile to become "frozen" in place by further growth. This rate, $r^* = \hat{k}e^{-G_{mc}} - \hat{k}e^{-2G_{se}}$, is related to the overall growth rate of the system [28]. As tiles that attach without any correct bonds ("doubly-mismatched" tiles) will detach very quickly, to first approximation, the only states that need to be considered are empty (E), correct tile (C), and "almost correct tile" (A)—a tile that is attached by one correct bond—along with frozen states for correct and almost correct tiles (FC and FA). These states are described in Fig. 1(d).

Numerous techniques have been studied to reduce such error rates, especially "proofreading" transformations that transform individual tiles into multiple tile blocks or sets of tiles [29,6,4,20]. These techniques have been shown to significantly reduce error rates both in simulation and experimentally [11,6,3]. Such techniques rely on changing tile systems at an abstract level, and reduce error rates of even ideal systems. However, in implementing the abstract logic of a tile system in actual DNA tiles, design complexities cause the system's kinetics to deviate from the default kTAM parameters. In particular, the single-stranded "sticky ends" that implement the abstract ends must be chosen from a finite sequence space to be both as uniform in binding energy and as orthogonal as possible. Deviations here can introduce further errors [10].

2 Theoretical Model

In the kTAM, G_{se} and G_{mc} are by default considered to be constant and independent of both tiles and sticky ends. A more detailed model cannot assume this. G_{mc} is dependent upon tile concentration: the value may be different for each tile type, and may change as free tiles are depleted by attachment. However, as experimental techniques exist to keep tile concentrations approximately constant throughout assembly [23], we will assume a time-invariant (but possibly tile type dependent) G_{mc}.

G_{se}, on the other hand, will depend upon the bonds between sticky ends. Ends with different sequences will have different free energies for binding to their complements, and some ends may be able to partially bind to ends that are only partially complementary (Fig. 1(a)). This results in a G_{se}^{ij} for each pair of sticky ends (i, j). In the default kTAM, all non-diagonal terms will be zero, and all diagonal terms will be equal. G_{se}^{ij} can thus be defined in terms of deviations from a reference G_{se}:

$$G_{se}^{ii} = G_{se} + \delta_i \qquad\qquad G_{se}^{ij} = s_{ij} G_{se} \text{ for } i \neq j \ . \tag{2}$$

Non-uniform sticky ends, with non-zero δ_i, will affect the detachment rate of correct and almost-correct tile attachments, while spurious non-orthogonal binding strengths s_{ij} will only decrease detachment rates for almost-correct and doubly-mismatched tile attachments. In the following theoretical analysis, the much lower likelihood doubly-mismatched interactions are ignored. For simulations, done with the Xgrow kTAM simulator [2], these interactions are taken into account when there is non-orthogonal binding.

2.1 Uniformity

Non-uniform sticky end energies have been simulated previously [10], but have not been studied analytically. In the kTAM, the growth rate of an assembly depends on the difference between on and off rates [28], which we approximate for a uniform system as $r^* = \hat{k}e^{-G_{mc}} - \hat{k}e^{-2G_{se}}$.

For a system with non-uniform energies, a tile attaching by two i bonds will have

$$r^* = \hat{k}e^{-G_{mc}} - \hat{k}e^{-2G_{se}-2\delta_i} = \hat{k}e^{-2G_{se}}\left(e^{\epsilon} - e^{-2\delta_i}\right)$$

where we define $\epsilon \equiv 2G_{se} - G_{mc}$, a measure of supersaturation: for an ideal system, $\epsilon = 0$ results in unbiased growth, whereas $\epsilon > 0$ results in forward growth and $\epsilon < 0$ causes crystals to shrink. As can be seen, the growth rate will depend on the δ_i's of the bonds in the growth region. With $\delta_i < -\frac{1}{2}\epsilon$ (negative δ corresponds to weaker binding), growth in a region won't be favorable.

In the worst case, where tiles attaching by two bonds with the smallest δ_i form a sufficiently large region, growth can only be ensured if $\epsilon > -2 \min\{\delta_i\}$, and error rates can be approximated by the kTAM with this minimum ϵ value. The kinetic trapping model in the default kTAM results in an error rate $P_{error} \approx me^{-G_{se}+\epsilon}$ for m possible incorrect tile attachments [28], so the worst-case error rate for a given $\delta_{\min} \equiv \min\{\delta_i\}$ would be

$$P_{\text{error}} \approx me^{-G_{se}-2\delta_{\min}} \ . \tag{3}$$

Fig. 2(a) shows simulations of the PM3 system with ϵ adjusted along the lines of our worst-case growth requirements. For positive deviations, where most ends remain at the same strength, assembly time is largely unchanged, while the error rate increases. For negative (weaker bond) deviations, where ϵ is adjusted, the error rate rises per Eq. 3, while the assembly time decreases sharply as most tiles attach with the same G_{se}^{ii} but are at a higher concentration.

While this method to adjust tile concentrations ensures crystal growth, it may not obtain the optimal trade-off between growth rate and error rate. This trade-off has been addressed for perfect sticky ends [5,12], but is more complicated with imperfect sticky ends and complex tile sets. Rather than simply adjusting all concentrations uniformly, the assumption can be made, which is not necessarily optimal, that error rates for a complex tile set can be reduced by ensuring

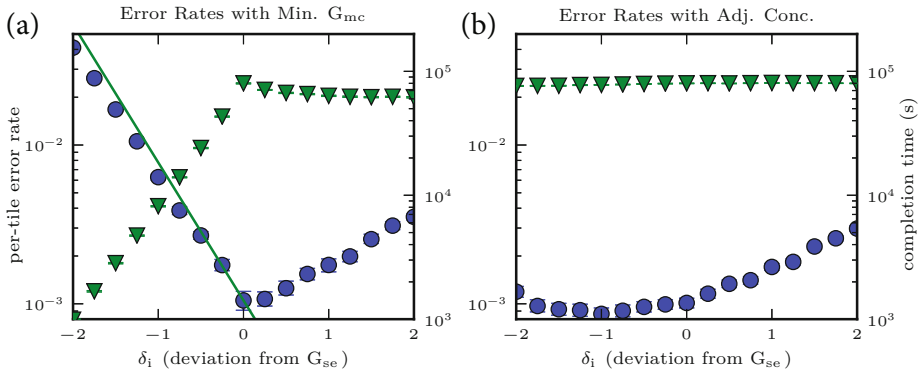

Fig. 2. Error rates for Pascal mod 3 systems with non-uniform end interactions simulated in Xgrow. In both (a) and (b), single sticky ends have been changed so that $G_{se}^{ii} = G_{se} + \delta_i$, while all others have remained at G_{se}. In (a), the ϵ for the system has been uniformly changed to always allow forward-growth by two of the weakest bond types by setting G_{mc}. In (b), the tiles with deviating ends have had their concentration adjusted so that all tiles have the same growth rate $r^* = \hat{k}e^{-G_{mc}^n} - \hat{k}e^{-G_{se}^{ii}-G_{se}^{jj}}$, where tile type n attaches using sticky end types i and j. Blue circles show error rates; green triangles show the time taken to construct an 8000 tile assembly, the line in (a) shows Eq. 3. For these simulations, we set base parameters of $G_{se} = 10$ and $G_{mc} = 19.2$.

that the overall growth rate remains uniform throughout the crystal. This can be achieved by modifying the concentrations of tiles to modify their G_{mc} values such that the r^* for each tile type is the same. Fig. 2(b) shows simulations of this form of concentration-adjustment with the PM3 system. As expected, assembly time remains almost completely unchanged across a large range of deviations. Meanwhile negative deviations do not significantly increase error rates, and positive deviations increase error rates in a manner similar to Fig. 2(a).

2.2 Orthogonality

Unlike non-uniformity, the kinetic trapping model for growth errors can be easily extended to account for non-orthogonality. Assuming $s_{ij} \ll 1$, growth errors will be primarily caused by almost-correct tiles attaching by one correct and one incorrect bond, as in the ideal case. A uniform incorrect bond strength of s, and m possible almost-correct tiles for a given lattice site, then gives the following rates of change between the different states shown in Fig. 1(d):

$$\dot{P}(t) = \begin{array}{c} \\ E \\ C \\ A \\ FC \\ FA \end{array} \begin{array}{c} \begin{array}{ccccc} E & C & A & FC & FA \end{array} \\ \begin{pmatrix} -2r_f & r_2 & r_{(1+s)} & 0 & 0 \\ r_f & -r_2 - r^* & 0 & 0 & 0 \\ mr_f & 0 & -r_{(1+s)} - r^* & 0 & 0 \\ 0 & r^* & 0 & 0 & 0 \\ 0 & 0 & r^* & 0 & 0 \end{pmatrix} \end{array} P(t) \quad . \tag{4}$$

Here $P(t)$ is a vector of probabilities at time t that the site will be in a state $[E, C, A, FC, FA]$. The steady state of this is not useful, as any combination of

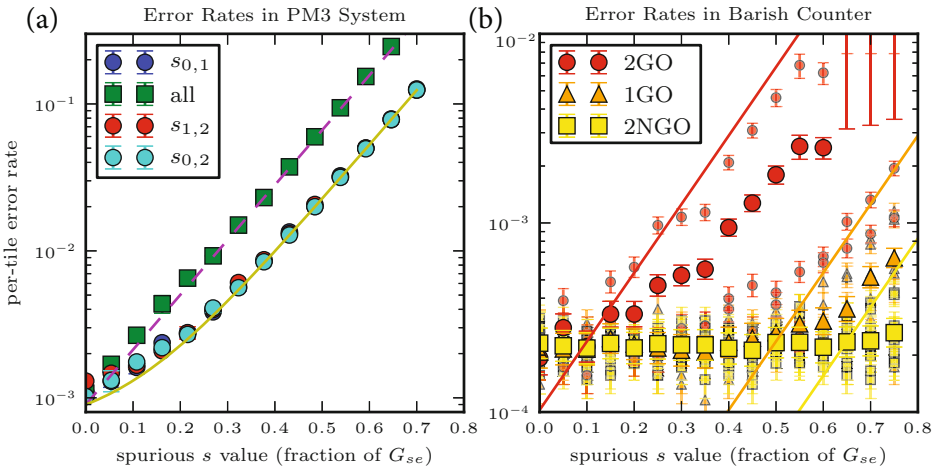

Fig. 3. Error rates with non-orthogonal interactions. (a) shows interactions for the PM3 system; circles and solid lines show simulated and theoretical error rates, respectively, with single pairs interacting. Squares and dashed lines show error rates for a uniform non-orthogonal interaction between every pair. (b) shows error rates for sensitive single non-orthogonal pairs in the Barish counter system, along with lines showing $e^{-(s-\sigma)G_{se}}$ for various values of σ chosen to roughly follow the worst pairs of each sensitivity. Small dots represent individual pairs, while large dots show averages for sensitivity classes. For (a) $G_{se} = 10$ and $G_{mc} = 19.2$, for (b) $G_{se} = 8.35$ and $G_{mc} = 17.8$.

FC and FA will be a steady state. Instead, the eventual probability of being in FA after starting only in state E at $t = 0$ will provide an error rate per additional tile in an assembly. This can be treated as a flow problem, where we consider the differential accumulation into FC and FA from E, as in Winfree [28]. From this, the probability of an almost-correct tile being trapped in place is:

$$P_{error} = \frac{m}{m + \frac{r_f + r_{1+s}}{r_f + r_2}} \approx \frac{1}{1 + \frac{1}{m} e^{(1-s)G_{se} - \epsilon}} \approx m e^{(s-1)G_{se} + \epsilon} \,. \tag{5}$$

While tile systems will have a different number of possible almost-correct tiles for different lattice sites, making this result less applicable, the PM3 system has an equal number for every possible lattice site. Fig. 3(a) shows error rates in simulations with interactions between single pairs of ends and for a uniform non-orthogonal interaction energy between every pair. In both cases, error rates largely follow Eq. 5.

2.3 Sticky End Sensitivity

When non-orthogonal sticky end interactions are not uniform, the degree of their influence on error rates may depend on which tile types they appear on and the logical interactions within the tile set. In systems where a tile never has

the opportunity to attach with strength $1 + s_{ij}$, interactions between i and j may be less relevant, whereas other pairs of ends in the system may allow tiles to erroneously attach during correct growth and be simply locked in place by continued growth. For example, Fig. 3(b) shows error rates for the Barish counter system when non-orthogonal interactions are introduced between single pairs of sticky ends. These pairs have been organized into sets (1NGO, 2NGO, 1GO, and 2GO) based on a model described below of how interactions between them may affect the tile system. As can be seen, this model has some success in predicting the impact different pairs will have on error rates.

We start by assuming that all attachments in growth occur with single tiles attaching by exactly two correct strength-1 bonds. Assuming that each tile in the system can have its ends labelled as inputs or outputs, and that every growth site has a unique tile that can attach by inputs, all lattice locations possible in the system will eventually be filled by a specific tile. Rather than looking at lattice sites that actually appear in correct growth, which would require simulation, we can combinatorially investigate all possible local neighborhoods that might appear, and conservatively examine them for possible problems. For example, whether there exists a tile that can attach with strength $1 + s_{ij}$ can be approximated by whether there are two tiles that share a common input bond on one side but not the other, so that when one tile incorrectly attaches where the other could attach correctly, it forms a strength 1 bond for the common bond and a strength s_{ij} bond for the mismatch (as in Fig. 4(a)).

We describe end pairs where such tiles exist as being in the set of "first-order sensitive" end pairs. If the sides of the tiles are inputs for at least one tile type, and thus the tiles can attach in normal forward growth, the end pair is in the set of first-order growth oriented sensitive (1GO) pairs, whereas without consideration of input and output sides, the end pair is in the set of first-order non-growth-oriented sensitive (1NGO) pairs. End pairs (i, j) that are in 1NGO but not 1GO have tiles that can attach with strength $1 + s_{ij}$ only during growth after an error or at sites where there is no correct tile.

While end pairs in these sets have tiles that allow the first, erroneous tile attachment in the kinetic trapping model, the model also requires that a second tile be able attach by two correct bonds to the erroneous tile and adjacent tiles to trap the error in place. This is also not necessarily possible: an incorrect attachment could result in there being no adjacent correct attachment, and designing systems where this is the case is in fact the goal of proofreading systems [29].

Thus we can devise "second-order sensitive" sets of end pairs that allow this second, correct tile attachment, and are therefore expected to be more likely to cause errors. Consider a pair of tiles A and X with a common bond on one side but not the other, satisfying the criteria for a first-order sensitive pair. Whether a further tile can attach with strength 2 can be approximated by whether there is some second pair of tiles, B and Y, that can each attach to some third side of their respective original tiles, and also share a common bond on another side. In a plausible local neighborhood where A and B could attach correctly in sequence, it is possible for X to first attach erroneously, with strength $1+s_{ij}$ (in the location

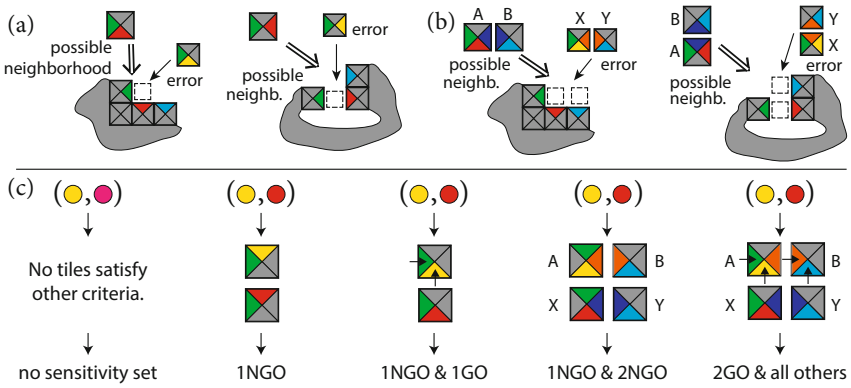

Fig. 4. Illustration of end pair sensitivity sets. For simplicity, all left and bottom sides are considered inputs. (a) shows, for given tiles, examples of possible local neighborhoods they could attach to and tiles that could erroneously attach via first-order sensitivity. (b) shows, for given pairs of tiles A and B, examples of local neighborhoods the pair could attach to in sequence, and a pair of tiles X and Y that could erroneously attach via second-order sensitivity. (c) shows examples of tiles satisfying various criteria for the shown end pairs to be in different sensitivity sets; arrows show examples of required input sides for growth-oriented sets.

where A could have bound), then for Y to attach with strength 2 (where B could have bound after A) owing to the second commond bond, as in Fig. 4(b).

As with first-order sensitivity, if the common and differing sides of the first pair of tiles are inputs, and sides of the second pair of tiles that are shared or attach to the first pair are also inputs, then the end pair involved is in the set of second-order growth oriented sensitive (2GO) pairs, whereas without consideration of inputs, the pair is in the set of second-order non-growth-oriented sensitive (2NGO) pairs.

These sets can be summarized more formally as follows, while examples of satisfying tiles are shown in Fig. 4(c):

- An end pair (i, j) is in the set of first-order sensitive end pairs if there exist at least two tiles in the tile system where both tiles share a common end k on one side, and on some other side, one tile has end i and the other has end j. If at least one of the two tiles has k and either i or j as inputs, then the end pair is in 1GO and 1NGO, otherwise, it is only in 1NGO.
- To determine if a first-order sensitive end pair (i, j) is in the set of second-order sensitive end pairs, consider a pair of tiles that satisfy the first-order criteria, and additional pairs of tiles that can attach to the first pair by bonds l and m (possibly the same) on a third side. If there exist a pair of these additional tiles that also share a common bond n, then the end pair is second-order sensitive. If at least one of the first tiles has k and either i or j as inputs, and one of the additional tiles attaching to it has n and either l or m as an input, then the end pair is in 2GO and 2NGO, otherwise, it is only in 2NGO.

Note that this analysis is done without determining what assemblies and thus what local neighborhoods actually form, so the combinations of inputs being considered might never appear during the growth of a correct assembly. As such, it is conceivable that, for example, an end pair could be in 2GO without ever having an effect in correct growth of an assembly. While this is a significant limitation, determining if a combination of inputs ever occurs, or if two tiles are ever assembled adjacent to each other, is in general undecidable by reduction to the Halting problem [27]. Furthermore, our current software treats all bonds as strength-1, and all tiles as single tiles, with double tiles being represented by a pair of single tiles with a fake bond that is then excluded from the sets; whilst the set definitions could be extended to account for double tiles and strength-2 bonds, we have not yet investigated the complexities involved.

Also, while pairs may be in either or both of 1GO or 2NGO, in all systems we have considered, all pairs in 1GO have also been in 2NGO, and there have been no pairs that are only in 1NGO. End pairs that aren't in *any* of these sets, and can be described as "zeroth-order," should have interactions between them that have a negligible effect on error rates in the kinetic trapping model.

Very rough theoretical estimates of the contributions that sensitive end pairs will have on a system can be obtained by considering the number of tiles that need to attach incorrectly. For pairs in 2GO, as only the initial tile will need to attach incorrectly before it can be locked in place by a correct attachment, the probablity of an error every time such a situation occurs is $\sim e^{(s-1)Gse}$. For those in 1GO but not 2GO, since there is no correct attachment after the first tile attaches incorrectly, at least one further incorrect attachment will be required, giving a probability of error $\sim e^{(s-2)Gse}$ or lower. For pairs only in 2NGO or 1NGO, the probability that the first tile can attach incorrectly will depend upon the likelihood that growth is proceeding in an incorrect direction, which in turn will depend upon numerous factors, but will usually require at least one previous incorrect attachment, giving another factor of $\sim e^{-Gse}$ on top of their GO counterparts.

For the Barish counter, there are 342 pairs of ends (helix direction prevents around half the ends from attaching to the other half). Of these, 22 are 2NGO, 9 are both 1GO and 2NGO, and 3 are also 2GO. Fig. 3(b) shows error rates for increasing values of s_{ij} where one pair has its value increased and all other spurious pairs are left with $s_{ij} = 0$. Each pair has been classified by its "worst" set. As can be seen, 2NGO pairs have little impact on error rates beyond those seen in the ideal kTAM, 1GO pairs start to have an effect after around $s_{ij} > 0.4$, and 2GO pairs are the most sensitive. In the case of the three 2GO pairs in the Barish counter, two cause errors that prevent correct growth in the next row without an additional error, explaining the significant difference between the most sensitive 2GO pair and the two less sensitive pairs.

3 Sequence Design and Assignment

3.1 Sequence Design

DNA sequence set design for molecular computation is a widely-studied problem. Different applications necessitate different constraints and approaches: longer sequences with less stringent requirements can be constrained with combinatorial methods like Hamming distance [13], while work on sequences with more stringent requirements have used thermodynamic constraints [25]. However, the basic goal shared throughout most of these algorithms is to find the largest set of DNA sequences that hybridize to their complements significantly better than to any other sequences in the set, or to find a set of a certain size with the best possible "quality"; in this the problem is similar to the maximum independent set problem, which is NP complete [7,18].

For sticky ends, the sequence lengths required, especially the 5 to 6 nt ends of DAO-E tiles, are shorter, and provide a smaller sequence space, than most other work has considered, with a few exceptions that have largely generated very small sets [25]. Using the end pair sensitivity model, we can reduce errors from non-orthogonal interactions by changing the assignment of sequences to abstract ends, as described later. However, we have no corresponding model to allow us to compensate for non-uniform energies.

The goal for our sequence design, therefore, is to find a requested number of sequences that (a) have non-orthogonal interactions less than a set constraint, and (b) have binding energies (melting temperatures) as uniform as possible given the orthogonality constraints. This contrasts with many sequence design algorithms, where a minimum melting temperature is of primary importance [24], and from algorithms that simply constrain melting temperatures to be within set constraints [25], in that our algorithm chooses a sequence with the *closest* melting temperature at each step.

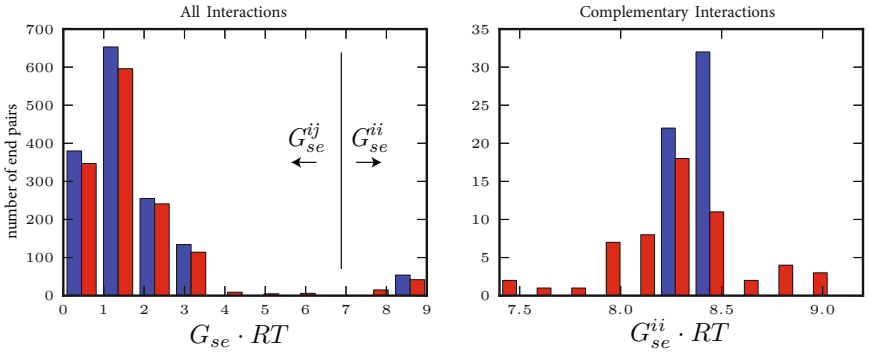

Fig. 5. Histograms of end pair interactions with the original Barish sequences (red) and newly designed sequences (blue). (a) shows all end pairs, (b) shows a zoomed-in area containing all end-complement pairs. All energies were calculated using the energy model in our sequence designer at 37 °C.

As the lengths of sticky end sequences are short, complex secondary structure is limited, and thus our algorithm uses an approximation of minimum free energy (MFE) for thermodynamic calculations. Similar to the "h-measure" used in Phan et al [18], the algorithm considers hybridization between two sequences with every possible offset, and uses the nearest-neighbor interaction data from SantaLucia et al [22], including values for symmetric loops, dangles, single-base mismatched pairs, and coaxial stacking with core sequences. Furthermore, for DAO-E tiles, core helix bases adjacent to the sticky ends affect energetics, and need to be designed alongside the sticky end sequences.

Our algorithm works as follows, for length L sticky ends.

1. Generate a set of all possible available sequences A that fit user requirements. With adjacent bases considered, this could be as many as 4^{L+2} sequences.
2. Calculate end-complement binding energies $G_{se}^{ii'}$ for all sequences in A, and (to speed up computation) remove any sequence that falls outside a user-specified range around the median $G_{se}^{ii'}$ of all sequences initially in A, which we call $\overline{G_{se}}$.
3. For each sequence needed:
   (a) Randomly choose a sequence i from all sequences in A that are closest to $\overline{G_{se}}$, and add this to the set of chosen sequences C.
   (b) Calculate the G_{se}^{ij} between i and every remaining sequence j in A, and remove all sequences from A with a G_{se}^{ij} greater than a user-specified value.
4. Stop when either A is empty, or a sufficient number of sequences have been generated.

$\overline{G_{se}}$ is chosen as the desired ideal G_{se} in order to ensure a large number of sequences with similar G_{se}^{ii}s will be available, for 5 nt ends, the desired value is $G_{se} \cdot RT = 8.35$ kcal/mol at $37°C$. By adjusting parameters, the maximum number of sequences that can be chosen can be changed as shown in Table 1; running the algorithm repeatedly will also find different numbers of sequences.

Sets chosen by this algorithm are guaranteed to have all ends interact less than a set amount $s_{ij} < s_{\text{desired}}$ with ends other than their complements, and to deviate from the desired correct interaction by less than a set amount $|\delta_i| < \delta_{\text{desired}}$, though when generating sets of a fixed size the largest δ_is will often be much smaller, as the software selects for the smallest δ_i values possible.

Fig. 5 shows a comparison between end pair interactions in the original Barish counter system and new sequences designed with our sequence design software. As can be seen, our software prevents large non-orthogonal interactions of 4 kcal/mol $< G_{se}^{ij} \cdot RT <$ 6 kcal/mol, but does not significantly reduce interactions with $G_{se}^{ij} \cdot RT <$ 4 kcal/mol. However, for complementary interactions, our software is able to find a significantly more uniform set of ends.

The practical value of this designer depends on the accuracy of the underlying energy model, of course, but the same algorithm can be used with different energy models as understanding of sticky end energetics is improved. The algorithm, with some energy model modifications, may also be of use in other

Table 1. Examples of the number of sticky ends found by our designer for varying user-specified parameters (bold). For lengths 5 and 6, examples are the best out of 100 runs, while for length 10, the example is a single run.

Length (nt)	$\overline{\boldsymbol{G_{se}}} \cdot \boldsymbol{RT}$	$\max(\boldsymbol{s_{ij}})$	# found	std(δ_i)	max δ_i
5	8.354	0.2	5	$0.04G_{se}$	$0.1G_{se}$
5	8.354	0.4	21	$0.01G_{se}$	$0.038G_{se}$
5	8.354	0.5	40	$0.01G_{se}$	$0.036G_{se}$
6	9.818	0.4	29	$0.004G_{se}$	$0.015G_{se}$
10	15.454	0.4	183	$0.01G_{se}$	$0.05G_{se}$

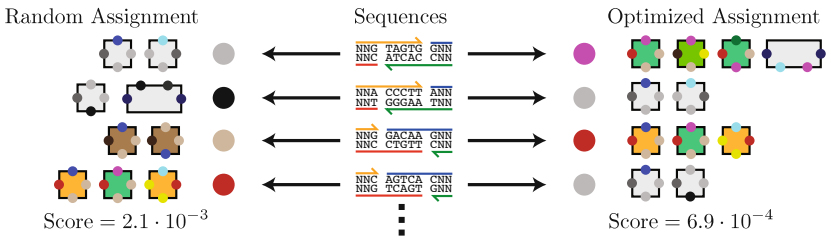

Fig. 6. Illustration of end assignment for the Barish counter set with new sequences. For conciseness, only a portion of the ends are shown.

areas of DNA computation where very short sequences with very similar melting temperatures and low non-orthogonal interactions are needed, such as toehold regions in strand displacement systems. However, it does not consider a number of factors important for actual strand displacement regions, and starts to become computationally intractable for sequences longer than 10 or 11 nt.

3.2 Sequence Assignment

The sequence designer is able to find sets of ends with very similar complementary interactions, and low non-orthogonal interactions. However, by ensuring that sequences are assigned to ends in a system such that end pairs with higher sensitivity have lower interactions, errors can further be reduced, and perhaps more importantly, the chance that a poor choice of sequences is made for a critical pair of ends can be minimized.

We assigned ends using a simulated annealing algorithm that used, as a score, the sum of rough error estimates for each end pair (see Fig. 4):

$$S(\text{assignment}) = \sum_{i,j \in 2GO} e^{-(s_{ij}-1.1)G_{se}} + \sum_{i,j \in 1GO \text{ and } \notin 2GO} e^{-(s_{ij}-1.5)G_{se}} + \sum_{i,j \in 2NGO \text{ and } \notin 1GO} e^{-(s_{ij}-1.65)G_{se}} + \sum_{i,j \in 1NGO \text{ and } \notin 2NGO} e^{-(s_{ij}-2)G_{se}} . \quad (6)$$

We call the resulting assignment 'optimized', although of course it is not guaranteed to be a global optimum. Offset values in the exponents were set

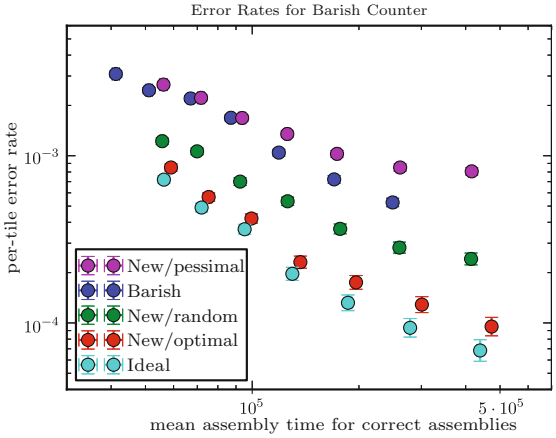

Fig. 7. Error rates for the Barish counter system with different sticky end sequences. Error rates are calculated from the percentage of correct assemblies formed of size 673. G_{se} values are calculated from ends, or are uniformly $G_{se} \cdot RT = 8.35$ kcal/mol in the ideal case. G_{mc} values were varied between 17.6 and 17.9. 1000 simulations were run for each G_{mc} value.

by rough estimates of the worst errors for different classes in the simulations shown in Fig. 4, and terms here for 2GO, 1GO and 2NGO are shown by solid lines in that figure. For 1NGO, the -2 parameter is chosen simply to be lower than other classes, as no system we have examined has end pairs that are only 1NGO. Since the sequence designer chooses adjacent bases as well as sticky end sequences, sequences can be consistently assigned to ends on all tiles, as in Fig. 6. The sequences and tiles for the Barish counter cannot be assigned in the same way, as different tiles with the same sticky end types often have different adjacent base pairs, modifying their interactions. Furthermore, as the sequence assignment algorithm only considers non-orthogonal interactions, results on a system with significant non-uniformity will likely be inconsistent.

Fig. 7 shows simulated error rates and assembly time for counters using sequences from Barish et al [3], sequences designed by our sequence designer and randomly assigned, and the same designed sequences assigned by our simulated annealing algorithm to both minimize and maximize the score in Eq. 6, along with error rates and assembly time for the system under ideal kTAM conditions. For a range of G_{mc} values and resultant assembly times, there is at least a 3-fold improvement in error rate between new sequences that are pessimally and optimally assigned by our scoring function, with increasing improvement as the assembly rate, and thus ideal error rate, decreases. For optimally assigned sequences, error remains close to the ideal error rate. The original sequences and assignment for the Barish counter perform slightly better than the pessimally assigned new sequences.

4 Conclusions and Discussion

These methods of sticky end design and assignment serve two purposes: firstly, to design experimental systems with error rates as close to the ideal kTAM as possible, and secondly, to reduce the chance that a poor choice of sequences, or even a poor assignment of sequences to tiles, might significantly impact experimental results. The methods should be relevant for most types of DNA tiles, and most tile systems with deterministic algorithmic behavior. Our software for these algorithms is available online [1].

The simulation results here, and the methods themselves, are reliant on the accuracy of the energy model used. While some research has been done on sticky-end energetics [15,19,14,9], usually for individual pairs of tiles, it is not known how well nearest-neighbor models of DNA energetics apply to sticky ends on DNA tiles in lattices. Different tile structures may also require slightly different models, especially with regard to coaxial stacking with base pairs adjacent to the sticky ends.

It is possible that extending end sensitivity definitions to higher orders, considering more than two tile attachments, may be a useful area of investigation, especially when considering tile systems making use of similarly higher order proofreading. Indeed, proofreading can counteract at a more fundamental level some of the same errors that arise from non-orthogonal interactions. The effects of non-uniform sticky end energies, however, may still significantly impact proofreading sets, and remain a potentially fruitful area of research beyond our simplistic modeling and concentration adjustment technique.

References

1. StickyDesign, `http://dna.caltech.edu/StickyDesign/`
2. The Xgrow simulator, `http://dna.caltech.edu/Xgrow/`
3. Barish, R.D., Schulman, R., Rothemund, P.W.K., Winfree, E.: An information-bearing seed for nucleating algorithmic self-assembly. Proc. Natl. Acad. Sci. USA 106, 6054–6059 (2009)
4. Chen, H.-L., Goel, A.: Error free self-assembly using error prone tiles. In: Ferretti, C., Mauri, G., Zandron, C. (eds.) DNA10. LNCS, vol. 3384, pp. 62–75. Springer, Heidelberg (2005)
5. Chen, H.-L., Kao, M.-Y.: Optimizing tile concentrations to minimize errors and time for DNA tile self-assembly systems. In: Sakakibara, Y., Mi, Y. (eds.) DNA16. LNCS, vol. 6518, pp. 13–24. Springer, Heidelberg (2011)
6. Chen, H.-L., Schulman, R., Goel, A., Winfree, E.: Reducing facet nucleation during algorithmic self-assembly. Nano Lett. 7, 2913–2919 (2007)
7. Deaton, R., Chen, J., Bi, H., Rose, J.: A software tool for generating non-crosshybridizing libraries of DNA oligonucleotides. In: Hagiya, M., Ohuchi, A. (eds.) DNA8. LNCS, vol. 2568, pp. 252–261. Springer, Heidelberg (2003)
8. Doty, D.: Theory of algorithmic self-assembly. Commun. ACM 55, 78–88 (2012)
9. Evans, C.G., Hariadi, R.F., Winfree, E.: Direct atomic force microscopy observation of DNA tile crystal growth at the single-molecule level. J. Am. Chem. Soc. 134, 10485–10492 (2012)

10. Fujibayashi, K., Murata, S.: Precise simulation model for DNA tile self-assembly. IEEE Trans. Nanotechnol. 8, 361–368 (2009)
11. Fujibayashi, K., Hariadi, R.F., Park, S.H., Winfree, E., Murata, S.: Toward reliable algorithmic self-assembly of DNA tiles: A fixed-width cellular automaton pattern. Nano Lett. 8, 1791–1797 (2008)
12. Jang, B., Kim, Y., Lombardi, F.: Error tolerance of DNA self-assembly by monomer concentration control. In: 2006 21st IEEE International Symposium on Defect and Fault Tolerance in VLSI Systems, pp. 89–97. IEEE (2006)
13. Kick, A., Bönsch, M., Mertig, M.: EGNAS: An exhaustive DNA sequence design algorithm. BMC Bioinformatics 13, 138 (2012)
14. Li, Z., Liu, M., Wang, L., Nangreave, J., Yan, H., Liu, Y.: Molecular behavior of DNA origami in higher-order self-assembly. J. Am. Chem. Soc. 132, 13545–13552 (2013)
15. Nangreave, J., Yan, H., Liu, Y.: Studies of thermal stability of multivalent DNA hybridization in a nanostructured system. Biophys. J. 97, 563–571 (2009)
16. Park, S.H., Yin, P., Liu, Y., Reif, J.H., LaBean, T.H., Yan, H.: Programmable DNA self-assemblies for nanoscale organization of ligands and proteins. Nano Lett. 5, 729–733 (2013)
17. Patitz, M.J.: An introduction to tile-based self-assembly. In: Durand-Lose, J., Jonoska, N. (eds.) UCNC 2012. LNCS, vol. 7445, pp. 34–62. Springer, Heidelberg (2012)
18. Phan, V., Garzon, M.H.: On codeword design in metric DNA spaces. Nat. Comput. 8, 571–588 (2008)
19. Pinheiro, A.V., Nangreave, J., Jiang, S., Yan, H., Liu, Y.: Steric crowding and the kinetics of DNA hybridization within a DNA nanostructure system. ACS Nano 6, 5521–5530 (2013)
20. Reif, J.H., Sahu, S., Yin, P.: Compact error-resilient computational DNA tiling assemblies. In: Ferretti, C., Mauri, G., Zandron, C. (eds.) DNA10. LNCS, vol. 3384, pp. 293–307. Springer, Heidelberg (2005)
21. Rothemund, P.W.K., Papadakis, N., Winfree, E.: Algorithmic self-assembly of DNA Sierpinski triangles. PLoS Biol. 2, e424 (2004)
22. SantaLucia, J., Hicks, D.: The Thermodynamics of DNA Structural Motifs. Annu. Rev. Biophys. Biomol. Struct. 33, 415–440 (2004)
23. Schulman, R., Yurke, B., Winfree, E.: Robust self-replication of combinatorial information via crystal growth and scission. Proc. Natl. Acad. Sci. USA 109, 6405–6410 (2012)
24. Tanaka, F.: Design of nucleic acid sequences for DNA computing based on a thermodynamic approach. Nucleic Acids Res. 33, 903–911 (2005)
25. Tulpan, D., Andronescu, M., Chang, S.B., Shortreed, M.R., Condon, A., Hoos, H.H., Smith, L.M.: Thermodynamically based DNA strand design. Nucleic Acids Res. 33, 4951–4964 (2005)
26. Wei, B., Dai, M., Yin, P.: Complex shapes self-assembled from single-stranded DNA tiles. Nature 485, 623–626 (2012)
27. Winfree, E.: On the Computational Power of DNA Annealing and Ligation. In: DNA Computers. DIMACS Series in Discrete Mathematics and Computer Science, pp. 199–221. AMS (1996)
28. Winfree, E.: Simulations of computing by self-assembly. Tech. Rep. Caltech CSTR:1998.22, California Inst. Technol., Pasadena, CA (1998)
29. Winfree, E., Bekbolatov, R.: Proofreading tile sets: Error correction for algorithmic self-assembly. In: Chen, J., Reif, J.H. (eds.) DNA9. LNCS, vol. 2943, pp. 126–144. Springer, Heidelberg (2004)