The Tile Complexity of Linear Assemblies

Harish Chandran, Nikhil Gopalkrishnan, and John Reif

Department of Computer Science, Duke University, Durham, NC 27707
{harish,nikhil,reif}@cs.duke.edu

Abstract. The conventional Tile Assembly Model (TAM) developed by Winfree using Wang tiles is a powerful, Turing-universal theoretical framework which models varied self-assembly processes. We describe a natural extension to TAM called the Probabilistic Tile Assembly Model (PTAM) to model the inherent probabilistic behavior in physically realized self-assembled systems. A particular challenge in DNA nanoscience is to form linear assemblies or *rulers* of a specified length using the smallest possible tile set. These rulers can then be used as components for construction of other complex structures. In TAM, a deterministic linear assembly of length N requires a tile set of cardinality at least N. In contrast, for any given N, we demonstrate linear assemblies of expected length N with a tile set of cardinality $\Theta(\log N)$ and prove a matching lower bound of $\Omega(\log N)$. We also propose a simple extension to PTAM called κ-pad systems in which we associate κ pads with each side of a tile, allowing abutting tiles to bind when at least one pair of corresponding pads match and prove analogous results. All our probabilistic constructions are free from co-operative tile binding errors and can be modified to produce assemblies whose probability distribution of lengths has arbitrarily small tail bounds dropping exponentially with a given multiplicative factor increase in number of tile types. Thus, for linear assembly systems, we have shown that randomization can be exploited to get large improvements in tile complexity at a small expense of precision in length.

1 Introduction

Biological systems show a remarkable range of form and function. How are these multitude of systems constructed? What are the principles that govern them? In particular, as computer scientists, we ask if there are simple rules whose repeated application can give rise to such complex systems. This leads us to the study of self-assembly.

1.1 Fundamental Nature of Self-assembly

Self-assembly is a fundamental pervasive natural phenomenon that gives rise to complex structures and functions. It describes processes in which a disordered system of pre-existing components form organized structures as a consequence of specific, local interactions among the components themselves, without any external direction. In its most complex form, self-assembly encompasses the processes involved in growth and reproduction of higher order life. A simpler example of self-assembly is the orderly growth of crystals. In the laboratory, self-assembly

S. Albers et al. (Eds.): ICALP 2009, Part I, LNCS 5555, pp. 235–253, 2009.

techniques have produced increasingly complex structures [1,2] and dynamical systems [3]. The roots of attempts to model and study self-assembly begin with the study of tilings.

A *Wang tile* [4] is an oriented unit square with a pad associated with each side. Any two tiles with the same pad on corresponding sides are said to be of the same *tile type*. Tile orientation is fixed, they cannot be rotated or reflected[1]. Given a finite set S of Wang tiles types, a valid arrangement of S on a planar unit square grid consists of copies of Wang tiles from the set S such that abutting pads of all pairs of neighboring tiles match. The *tiling* or *domino* problem for a set of Wang tiles is: can tiles from S (chosen with replacement) be arranged to cover the entire planar grid? Berger [5] proved the undecidability of the tiling problem by reducing the halting problem [6] to it. Robinson [7] gave an alternative proof involving a simulation of any single tape deterministic Turing Machine by some set of Wang tiles. Garey and Johnson [8] and Lewis and Papadimitrou [9] proved that the problem of tiling a finite rectangle is NP-complete. These results paved the way for Wang tiling systems to be used for computation. But, Wang tilings do not model coordinated growth and hence do not describe complex self-assembly processes. Winfree [10] extended Wang tilings to the Tile Assembly Model (TAM) with a view to model self-assembly processes, laying a theoretical foundation [11,12] for a form of DNA based computation, in particular, molecular computation via assembly of DNA lattices with tiles in the form of DNA motifs.

The *tile complexity* [13] of assembling a shape is defined as the minimum number of tile types for assembling that shape. Tile complexity, apart from capturing the information complexity of shapes, is also important as there exist fundamental limits on the number of tile types one can design using DNA sequences of fixed length. Various ingenious constructions for shapes like squares [14], rectangles and computations like counting [15], XOR [16] etc. exist in this model. Lower bounds on tile complexity have also been shown for various shapes. Stochastic processes play a major role in self-assembly and have been investigated theoretically by Winfree [17] and Adleman [11] and in the laboratory by Schulman et al. [18]. However, TAM is deterministic in the sense that it produces exactly one terminal assembly given a tile set. This is because at most one type of tile is allowed to attach at any position in a partially formed assembly. See Section 2 for more details. This work investigates the effects of relaxing these constraints and reduces the number of tile types required to form linear assemblies of given length. In contrast to earlier work in stochastic self-assembly, we make tile attachments irreversible (as in TAM) and allow multiple tile types to attach at any position.

1.2 Motivation

A particular challenge in DNA nanoscience is to form linear assemblies or *rulers* of a specified length from unit sized square tiles. These rulers can then be used

[1] This is a valid assumption when implementing Wang tiles in the laboratory using DNA due to the complimentary nature of DNA strand binding.

as a component for construction of other complex structures. One can use these structures as beams and struts within the nanoscale (See Fig.1a). Linear assemblies can also serve as boundaries [18] and nucleation sites for more complex nanostructures. (Note that due to the inherent flexible nature of linear nanostructures, most complex nanostructures will generally tolerate small deviations from the intended lengths of these substructures). Various tile based techniques for constructing linear assemblies have been successfully explored in the laboratory [19,18]. Hence, tile assembly models for linear assemblies are apt theoretical frameworks for exploring a fundamental and important challenge in DNA nanoscience. In TAM, rulers of length N can be trivially constructed by deterministic assembly of N distinct tile types. This is also the matching linear lower bound for size of tile sets in deterministic TAM, as shown in Section 4. Thus, it is impractical to form large linear structures using the deterministic techniques of TAM. Long thin rectangles (which are approximations of linear assemblies) can be formed using $\Theta(\frac{\log N}{\log\log N})$ tile types but they suffer errors due to co-operative tile binding. In contrast with linear assemblies, the number of tile types to form an $N \times N$ square is only $\Theta(\frac{\log N}{\log\log N})$ [14], which is exponentially better than the lower bound for linear assemblies. This bound for squares is asymptotically tight for almost all N as dictated by information theory[13] while the one for linear assemblies is not. This begs the question: why are we not able to reach information theoretic limit of $\Theta(\frac{\log N}{\log\log N})$ in linear structures using TAM? Is this lower bound tight? What is the longest (finite) linear assembly one can assemble with a set of n tile types in realistic tiling models? What changes to TAM will give us the power to specify the linear systems using a smaller tile set? While square assemblies have been extensively studied [13,14,20,21], many questions remain about linear assemblies, which are simpler constructs yet are fundamental building blocks at the nanoscale. We answer a number of these questions and show novel, interesting results using techniques that differ considerably from existing ones. While there have been numerous variations on TAM in recent years, their impact on laboratory techniques in DNA self-assembly are minimal. At the same time, design principles used in DNA self-assembly do not fully leverage the programmability and stochasticity inherent to self-assembly. Hence, our goal is to develop a simple model that directs design principles of experimental DNA self-assembly by taking advantage of inherent stochasticity of self-assembly. It is noteworthy that the techniques for designing and analyzing these simple constructs under our simple model are non-trivial and theoretically rich.

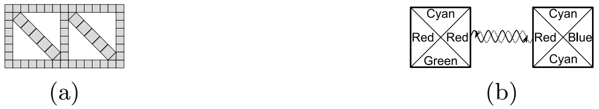


Fig. 1. (a) Possible nanostructures using rulers as substructures. (b) Diagonal tiles: Colors indicate pad type. Red pads are implemented using complimentary DNA. Strands for other pads are omitted.

1.3 Related Work in Self-assembly Using Probabilistic and Randomized Models

Non-determinism was used in tiling by Lagoudakis et al. [22] for implementing an algorithm for SAT. Recently, Becker et al. [23] describe probabilistic tile systems that yield squares, rectangles and diamond in expectation using $O(1)$ tile types. This work was extended by Ming-Yang Kao et al. [21] to yield arbitrarily close approximations to squares with arbitrarily high probability using $O(1)$ tile types. Both these papers allow precise arbitrary relative concentrations of tile types with no cost incurred in tile complexity. In the laboratory, achieving precise arbitrary relative concentrations between tiles is infeasible. Also, the descriptional complexity of tile systems in such models include not just the descriptional complexity of the tile set, but also the descriptional complexity of the concentration function. Thus, size of tile set producing an assembly is not a true indicator of its complexity. In PTAM, the set of tiles is a multi-set that implicity defines relative concentrations and precludes arbitrary relative concentrations. Thus, size of the tile set producing an assembly is a true indicator of its complexity. In addition, all our constructions have equimolar tile concentration and hence are experimentally feasible. Reif [24], and later Demaine et al. [25] discuss *staged self-assembly.* Demaine et al. [25] show how to get various shapes using $O(1)$ pad types. Aggarwal et al. [20] introduce various extensions to TAM and study the impact of these extension on both running time and the number of tile types. Compared to the above, PTAM is a simple extension to TAM that requires no laboratory techniques beyond those used to implement TAM. In particular, we consider standard one pot reaction mixtures with no intermediate purification steps. The Kinetic Tile Assembly Model (kTAM) proposed by Winfree [17] models kinetics and thermodynamics of DNA hybridization reactions. Schulman et al. [18] used DX tiles consisting of DNA stands to create one dimensional boundaries within the nanoscale. Adleman [11] proposed a mathematical theory of self-assembly which is used to investigate linear assemblies. While many fundamental theoretical questions arise in these models, the question of tile complexity of linear assemblies is uninteresting due the existence of the trivial lower bound mentioned in Section 1.2. Thus, the questions about linear self-assemblies examined in this paper are original and the constructions presented are novel.

1.4 Main Results

We describe a natural extension to TAM in Section 3 to allow randomized assembly, called the Probabilistic Tile Assembly Model (PTAM). A restriction of the model to diagonal, haltable, uni-seeded, and east-growing systems (defined in Section 3), which we call the standard PTAM is considered in this paper. Prior work in DNA self-assembly strongly suggests that standard PTAM can be realized in the laboratory. We show various non-trivial probabilistic constructions in PTAM for forming linear assemblies with a small tile set in Section 4, using techniques that differ considerably from existing assembly techniques. In particular, for any given N, we demonstrate linear assemblies of expected length N with tile set of cardinality $\Theta(\log N)$ using one pad per side of each tile in Section 4.2. We derive a matching lower bound of $\Omega(\log N)$ on the tile complexity of linear assemblies of any given expected length N in standard PTAM systems using one pad per side of each tile in Section 4.3. This lower bound, which holds for all N, is tight and better than the information theoretic lower bound of $\Omega(\frac{\log N}{\log \log N})$ which holds only for almost all N. We also propose a simple extension to PTAM in Section 5 called κ-pad systems in which we associate κ pads with each side of a tile, allowing abutting tiles to bind when at least one pair of corresponding pads match. This gives linear assemblies of expected length N with 2-pad (two pads per side of each tile) tile set of cardinality $\Theta(\frac{\log N}{\log \log N})$ tile types for infinitely many N. We show that we cannot achieve smaller tile complexity by proving a lower bound of $\Omega(\frac{\log N}{\log \log N})$ for each N on the cardinality of the κ-pad (κ pads per side of each tile) tile set required to form linear assemblies of expected length N in standard κ-pad PTAM systems for any constant κ. The techniques used for deriving these lower bounds are notable as they are stronger and differ from traditional Kolmogorov complexity based information theoretic methods used for lower bounds on tile complexity. Kolmogorov complexity based lower bounds do not preclude the possibility of achieving assemblies of very small tile multiset cardinality for infinitely many N while our lower bounds do, as they hold for every N. All our probabilistic constructions can be modified to produce assemblies whose probability distribution of lengths has arbitrarily small tail bounds dropping exponentially with a given k at the cost of a multiplicative factor k increase in number of tile types, as proved in Section 6.

2 The Tile Assembly Model for Linear Assemblies

This section describes the Tile Assembly Model (TAM) by Winfree for linear (1D) assemblies (henceforth referred to as LTAM). For a complete and formal description of the model see [13]. LTAM describes deterministic linear assemblies. The next section extends the model by introducing randomization. This paper considers only one-dimensional grid of integers $\mathbb{Z}$ which simplifies the definitions of the model. The directions $\mathfrak{D} = \{\text{East}, \text{West}\}$ are functions from $\mathbb{Z}$ to $\mathbb{Z}$, with $\text{East}(x) = x + 1$ and $\text{West}(x) = x - 1$. We say that x and x' are neighbors if $x' \in \{\text{West}(x), \text{East}(x)\}$. Note that $\text{East}^{-1} = \text{West}$ and vice versa. $\mathbb{N}$ is the set of natural numbers.

A *Wang tile* over the finite set of distinct *pads* Σ is a unit square where two opposite sides have pads from the set Σ^2. Formally, a tile t is an ordered pair of pads $(W_t, E_t) \in \Sigma^2$ indicating pad types on the West and East sides respectively. Thus, a tile cannot be reflected. For each tile t, we define $\text{pad}_{\text{East}}(t) = E_t$ and $\text{pad}_{\text{West}}(t) = W_t$. Σ contains a special *null pad*, denoted by ϕ. The *empty* tile (ϕ, ϕ) represents the absence of any tile. Pads determine when two tiles attach.

[2] In general, for two dimensional assemblies, tiles have pads on all four sides. However, we do not use any pads on the North and South sides in this paper and hence omit them. Also, we allow for multiple pads on the sides of a tile in Section 5.

A function $g : \Sigma \times \Sigma \rightarrow \{0, 1\}$ is a binary *pad strength function* if it satisfies $\forall x, y \in \Sigma,\ g(x, y) = g(y, x)$ and $g(\phi, x) = 0$. Linear assemblies do not have co-operative tile binding, i.e, interactions of more than one pair of pads at a given step. Hence the temperature parameter used in TAM is redundant in linear assemblies where tiles have only one pad per side. Throughout this paper we assume only a binary pad strength function. In this model each tile has only a single pad on each of its sides (West and East) whereas in Section 5 we allow multiple pads per side for each tile.

A *linear tiling system*, $\mathbb{T}$, is a tuple $\langle T, S, g \rangle$ where T containing the empty tile is the finite set of tile types, $S \subset T$ is the set of *seed* tiles and g is the binary pad strength function. A *configuration* of T is a function $A : \mathbb{Z} \rightarrow T$ with $A(0) = s$ for some $s \in S$. For $D \in \mathfrak{D}$ we say the tiles at x and $D(x)$ *attach* if $g(\text{pad}_D(A(x)), \text{pad}_{D^{-1}}(A(D(x)))) = 1$. *Self-assembly* is defined by a relation between configurations, $A \rightarrow B$, if there exists a tile $t \in T$, a direction $D \in \mathfrak{D}$ and an empty position x such that t attaches to $A(D(x))$. We define $A \xrightarrow{*} B$ as the reflexive transitive closure of $\rightarrow$ and say B is *derived* from A. For all $s \in S$ a *start* configuration start_s is given by $\text{start}_s(0) = s$ and $\forall x \neq 0 : \text{start}(x) = \text{empty}$. A configuration B is *produced* if $\text{start}_s \xrightarrow{*} B$ for some $s \in S$. A configuration is *terminal* if it is produced from start_s for some $s \in S$ and no other configuration can be derived from it. $\text{Term}(\mathbb{T})$ is the set of terminal configurations of $\mathbb{T}$. In TAM, a terminal configuration is thought of as the output of a tiling system given a seed tile $s \in S$. TAM requires that there be a unique terminal configuration for each seed. Note that it allows different attachment orders as long as they produce the same terminal configuration. This unique terminal configuration requirement means that given any non terminal configuration A, at most one $t \in T$ can attach at any given position. In this sense, TAM is deterministic. In the next section we will explore the effect of relaxing this condition of TAM.

DNA nanostructures can physically realize TAM as shown by Winfree et al. [10] with the DX tile and LaBean et al. [26] with the TX tile. Like the square tile in TAM, the DX and TX have *pads* that specify their interaction with other tiles. The pads are DNA sequences that attach via hybridization of complimentary nucleotides. Mao et al. [27] performed a laboratory demonstration of computation via tile assembly using TX tiles. Yan et al. [16] performed parallel XOR computation in the test-tube using Winfree's DX tile. Other simple computations have also been demonstrated. However, large and more complex computations are beset by errors and error correction remains a challenge towards general computing using DNA tiles.

3 The Probabilistic Tile Assembly Model

In TAM, the output of a tile system is said to be a shape of given fixed size (for example, square of side N, linear assemblies of length N) if the tile system *uniquely* produces it. In this paper, we consider some implications of relaxing this requirement. Instead of asking that a set of tiles produce a *unique* shape, we allow the set of terminal assemblies to contain *more than one shape* by designing tile systems which admit multiple tile type attachment at a given position in a configuration. Note that we do not allow pad mismatch errors. We also associate a probability of formation with each terminal assembly. These extensions and modifications to TAM are formalized for linear assemblies. Note that the definitions given below can be easily extended to assemblies in two-dimensions by introducing pads on North and South sides of tiles and including a temperature parameter τ as in [13] for co-operative binding effects.

3.1 The Probabilistic Tile Assembly Model (PTAM) for Linear Assemblies

A *probabilistic linear tiling system* $\mathbb{T}$ is given by the tuple $\langle T, S, g \rangle$, where T is a (finite) multiset of tile types, $S \subset T$ is the multiset of seed tiles and g is the binary pad strength function. The set of pad types Σ, tiles and configurations for $\mathbb{T}$ are defined as in Section 2. The *multiplicity* $\mathcal{M} : \Sigma \times \Sigma \rightarrow \mathbb{N}$ of a tile type is the number of times it occurs in T. T contains the empty tile type with $\mathcal{M}(empty) = 1$. Multiplicity models concentration. We assume a well-mixed, one pot reaction environment in which at each step some member of T is copied (chosen with replacement) from the pot with uniform probability. If the tile thus obtained can attach to the produced configuration, it does so, else a new member of T is copied with uniform probability in the next step. This continues till either a match is found or none exists, in which case the system halts. Note that this is a Gillespie simulation [28] with a seed serving as a nucleation site. A system with only one seed, $S = \{s\}$, is called *uni-seeded.* We consider only uni-seeded systems in this paper. The function *type(t)*, type : $T \rightarrow \Sigma \times \Sigma$, returns the tile type for any $t \in T$.

Self-assembly of a linear tiling system $\mathbb{T}$ is defined by a relation between set of positive probabilities and pair of configurations A and B as: $A \rightarrow^{p}_{\mathbb{T}} B$ (read as A gives B with probability p) if there exists a tile $t \in T$, a direction $D \in \mathcal{D}$ and an empty position x such that t attaches to $A(D(x))$ with positive probability p to give B where $p = \mathcal{M}(\text{type}(t))/\sum_{j \in \Delta} \mathcal{M}(\text{type}(j))$ where $\Delta = \{j | \text{ type}(j) \text{ attaches to } A(D(x))\}$. The closure of $\rightarrow^{p}_{\mathbb{T}}$, denoted by $\overset{*}{\rightarrow}{}^{\hat{p}}_{\mathbb{T}}$ (read as 'derives'), is defined by the following transitive law: if $A \rightarrow^{p_1}_{\mathbb{T}} B$ and $B \rightarrow^{p_2}_{\mathbb{T}} C$ then $A \rightarrow^{p_1 p_2}_{\mathbb{T}} C$. A configuration B is *produced* with positive probability p if $\text{start}_s \overset{*}{\rightarrow}{}^{p}_{\mathbb{T}} B$. A configuration is *terminal* if it is produced from start_s and no other configuration can be derived from it with positive probability. Term($\mathbb{T}$) is the set of terminal configurations of $\mathbb{T}$. We associate a *probability of formation*, $P(A)$ to each produced configuration A recursively, as follows: $P(\text{start}_s) = 1$ and $P(B) = \sum_{\Gamma} p_k P(A_k)$ where $\Gamma = \{k | A_k \rightarrow^{p_k}_{\mathbb{T}} B\}$. *Length* of a produced configuration A, written as $|A|$, is the number of non-empty tiles in it.

A configuration A is called a *linear assembly of length N* if it is terminal and $|A| = N$. Following Rothemund and Winfree's terminology [13], a linear tiling system is defined to be *diagonal* iff $g(x, y) = 0$ for all x, y with $x \neq y$ and $g(x, x) = 1$ for all $x \neq \phi$. A tile t is *reachable* in $\mathbb{T}$ if it is part of some produced configuration. A tile $t \in T$ is a *capping tile* if t is reachable and there exists $D \in \mathfrak{D}$ such that $g(\text{pad}_D(t), \text{pad}_{D^{-1}}(t')) = 0$ for each $t' \in T$. For $D =$ East the tile is called *East capping* and for $D =$ West it is called *West capping.* A capping tile halts growth in either the East or West direction. Note that a tile other than the seed cannot be both East and West capping. A linear probabilistic tiling system $\mathbb{T}$ is *haltable* iff for each produced configuration A, there exists a terminal configuration B such that $A \overset{*\ p}{\rightarrow}_{\mathbb{T}} B$ with positive probability p. Each terminal configuration has a probability of formation associated with it. If $\mathbb{T}$ is haltable, some terminal configuration occurs with certainty as stated without proof in the following Lemma.

Lemma 1. *If $\mathbb{T}$ is a haltable probabilistic linear tiling system, then $\sum_{A \in Term(\mathbb{T})} P(A) = 1$.*

A linear tiling system is called *east-growing* if the West pad of the seed tile is ϕ. A *simulation* of a probabilistic tile system $\mathbb{T}$ by a probabilistic tile system $\mathbb{Q}$ is a bijection f between terminal configurations that preserves lengths and probabilities of formation of assemblies, i.e. $f : \text{Term}(\mathbb{T}) \rightarrow \text{Term}(\mathbb{Q})$ satisfying $|A| = |f(A)|$ and $P(A) = P(f(A))$ for each $A \in \text{Term}(\mathbb{T})$. Any probabilistic linear tiling system $\mathbb{T}$ can be simulated by an east-growing probabilistic linear tiling system $\mathbb{Q}$ using no more than twice the number of tile types of $\mathbb{T}$, in the following manner. For the seed $s = (W_s, E_s)$ of $\mathbb{T}$, let $s' = (\phi, E'_s)$ be the seed of $\mathbb{Q}$ and for each East-capping tile $c = (W_c, E_c)$ of $\mathbb{T}$ let $\mathbb{Q}$ contain tile $c' = (W'_c, W''_s)$. For all other tiles $t = (W_t, E_t)$ of $\mathbb{T}$, let $\mathbb{Q}$ contain tiles $t_r = (W'_t, E'_t)$ and $t_l = (E''_t, W''_t)$. The reader may verify that this is a simulation. Hence, we consider only east-growing tile systems in this paper. A probabilistic linear tiling system is *equimolar* if $\forall t \in T : \mathcal{M}(t) = 1$. Thus, for an equimolar tile system, the cardinality of T equals the number of tile types in it. A probabilistic linear tiling system is *two-way branching* if at most two tile types can attach at any given position for any given configuration. A probabilistic linear tiling system is *standard* if it is diagonal, haltable, uni-seeded and east-growing.

Diagonal tile systems were suggested by Winfree and Rothemund [13]. These systems are implementable using DNA tiles. Matching pads are implemented as perfect Watson-Crick complimentary DNA sequences (see Fig.1b). Non-diagonal tile systems are not implementable using this technique. For tile systems producing linear assemblies that are not haltable, the expected length of the assembly diverges. For linear assemblies, no advantage in tile complexity or tail bounds on length of assemblies results from using multiple seeds. *Thus, we consider only standard systems in this paper.* Achieving arbitrary concentration vectors is infeasible in laboratory implementations using molecules. In contrast, equimolar systems are frequently achieved by chemists for various reactions. We demonstrate a equimolar standard linear tiling system whose tile complexity matches the more general lower bound of $\Omega(\log N)$ applicable to all standard linear tiling systems.

3.2 Complexity Measures for Tile Systems

Recall that the tile complexity of a shape in TAM is defined as the number of different tile types in the smallest tile set that realizes the shape. The tile complexity in TAM is closely related to the size of the smallest Turing machine describing the shape [29]. While in TAM the shape is realized deterministically, in PTAM we drop the requirement that a shape be obtained uniquely and instead ask that it be approximated by our probabilistic tile systems. What should be the correct measure of descriptional complexity of shapes in such probabilistic systems? Consider a probabilistic linear tiling system with with three tile types (Seed, Growth and Halt) at a $1 : N : 1$ relative concentration such that it assembles into a linear assembly of expected length $N + 2$. Clearly, the number of distinct tile types does not completely describe the assembly process in the absence of information about relative concentrations. Thus concentrations must be taken into account in any measure that hopes to intrinsically capture descriptional complexity. There exist modifications of TAM [21,25,20] where the number of tile types does not correspond to the descriptional complexity of the shape. These systems encode the complexity elsewhere, like in the concentration, temperature, mechanism etc. In contrast, *the standard systems of* PTAM *encode all the description of the shape in the tile multiset* through multiplicity of a tile type which models its concentration. Thus, the (probabilistic) descriptional complexity of shapes corresponds to the cardinality of the tile multiset which we call tile complexity. Note that multiplicity of tiles in the multiset count distinctly towards tile complexity.

What is the effect of the probabilistic model on tile complexity? We demonstrate linear assemblies of fixed expected length N using a tile set of small cardinality. In general, we are asking if there is any benefit in sacrificing the exact description of a shape for a probabilistic description. For linear assemblies, the answer is yes, as we show in the next section.

4 Constructing Linear Assemblies of Expected Length N

In the standard TAM, the tile complexity for a linear assembly of length N is N. This is because if a tile type occurs at more than one position in the assembly, the sub-unit between these two positions can repeat infinitely often. This does not produce a linear assembly of length N. The PTAM does not suffer from this drawback. By making longer and longer chains less likely, we ensure that most chains are of length close to N. All our constructions can be shown to have exponentially decaying tail with a linear multiplicative increase in the number of tile types. So we focus on the expected lengths of linear assemblies in the following sections. All of our constructions for linear assemblies of expected length $N \in \mathbb{N}$ are standard, equimolar and two-way branching. The random variable L always denotes the length of the assembly. Specific tiles systems in the rest of this section are illustrated using *tile binding diagrams*. Each tile type is represented by a square, with labels distinguishing different tile types. All possible interactions among tiles are denoted via arrows that originate at the West side of some tile and terminate on the East side of some tile, indicating pad strengths of 1 between these tiles along these sides. Absence of arrows indicate that no possible attachment can occur, i.e. pad strength is 0. Thus, all our systems are temperature 1 assemblies which are more resilient to errors than assemblies at greater temperatures. The latter suffer errors due to co-operative tile binding [30,31]. Moreover, temperature 1 systems are easier to implement in the laboratory than higher temperature systems. Since we consider only equimolar systems for the rest of this section, the cardinality of our tile multisets equal the number of tile types. We use these terms interchangeably for equimolar systems.

4.1 Linear Assemblies of Expected Length N Using $O(\log^2 N)$ Tile Types

In this section we present a standard linear tiling system that achieves a linear assembly of expected length N for any given N using $O(\log^2 N)$ tile types. First, we give a construction for powers of two, i.e. for any given $N = 2^i$ for some $i \in \mathbb{N}$, we show how to construct N length linear assemblies using $\Theta(\log N)$ tile types. Then we extend this construction to all N by expressing N in binary and linking together the chains corresponding to 1s in the binary representation of N.

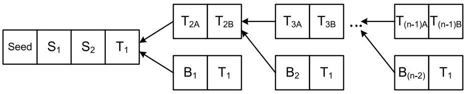


Fig. 2. Tile Binding Diagram for Powers of Two Construction

Powers of Two Construction: Fig.2 illustrates the tile set of size $3n + 2 = \Theta(n)$, used in a powers of two construction. The assembly halts only when the sequence $T_1, T_{2A}, T_{2B}, \ldots T_{(n-1)A}, T_{(n-1)B}$ of attachments is achieved. The bridge tiles B_i, $i = 1, 2 \ldots, n-2$, act as reset tiles at each stage of the assembly. Each probabilistic choice is between a reset in the form of B_i and progress towards completion in the form of $T_{(i+1)A}$. Attachment of T_1 to B_i and of T_{iB} to T_{iA} is deterministic.

Lemma 2. *Let L be the random variable equal to the length of the assembly. Then, $E[L] = 2^n$. Thus, an assembly of expected length 2^n can be constructed using $\Theta(n)$ tile types for any given $n \in \mathbb{N}$.*

Proof. We associate a sequence of independent Bernoulli trials, say coin flips, with the assembly process. Let the addition of the $\langle B_i, T_1 \rangle$ complex correspond to *Tails* and the addition of the $\langle T_{iA}, T_{iB} \rangle$ complex correspond to *Heads.* Halting of the assembly then corresponds to achieving a sequence of $n-2$ successive heads, corresponding to the sequence $\langle T_{2A}, T_{2B} \rangle, \ldots \langle T_{(n-1)A}, T_{(n-1)B} \rangle$ of attachments. The expected number of fair coin tosses for this to happen is $2(2^{(n-2)} - 1)$ [32]. Each coin toss adds two tiles to the linear assembly. Hence $E[L] = 4 + 4(2^{(n-2)} - 1) = 2^n$.

Extension to Arbitrary N: We extend the powers of two construction to all N by expressing N in binary, denoted by $B(N)$. For each i^{th} bit of $B(N)$ $(i > 2)$ equal to 1, we have a power of two construction of expected length 2^i, using $3i - 2$ tile types as in 4.1. We simply append these various constructions deterministically, and rely on linearity of expectation to achieve a linear assembly of length N in expectation.

Theorem 1. *An assembly of expected length N can be constructed using $O(\log^2 N)$ tile types for any given $N \in \mathbb{N}$.*

4.2 Linear Assemblies of Expected Length N Using $\Theta(\log N)$ Tile Types

In this section we present a standard linear tiling system that achieves linear assembly of length N in expectation for any given N using $\Theta(\log N)$ tile types. For powers of two, this construction reduces to one similar to that in Section 4.1. Our construction for general N is a more succinct than the one presented in Section 4.1. This new construction rests on the observation that the expected number of tiles of each type present in the powers of two construction decrease geometrically.

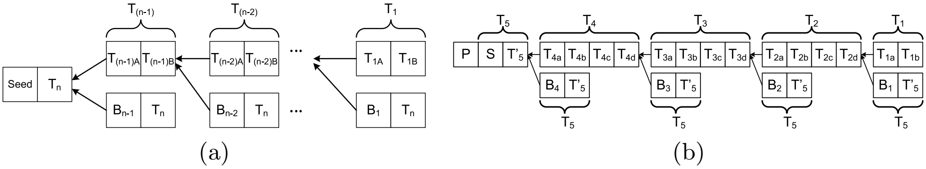


Fig. 3. Tile binding diagrams for $O(\log N)$ construction. (a) Tile Binding Diagram for Section 4.2. (b) Tile Binding Diagram: $N = 91; N'' = 90; N' = \frac{N''}{2} = 45 = (12221)_{\text{alt2}}$. P is the prefix tile.

Consider the linear tiling system depicted in Fig.3a. The size of the tile set is $3n - 1 = \Theta(n)$. The expected length of the assembly is $N = 2^{n+1} - 2 = \Theta(2^n)$ [32]. We observe that the number of *bi-tiles* [3] of type T_i decrease geometrically as i decreases as stated below.

Lemma 3. *Let X_i be the random variable equal to the number of bi-tiles of type T_i in the final assembly. Then $E[X_{i-1}] = \frac{E[X_i]}{2}$ and hence $E[X_i] = 2^{i-1}$ for $i = 2, 3, \ldots, n$*

Proof. Every time a bi-tile of type T_i appears, a bi-tile of type T_{i-1} follows immediately in the resulting assembly with probability 1/2 for $i = 2, 3 \ldots n$. So $E[X_{i-1}] = \frac{E[X_i]}{2}$. This property allows us to calculate the expected number of bi-tiles of each type. T_1 is the terminal bi-tile and appears exactly once. Hence its expectation is $1 = 2^0$. Repeated application of the above geometric decrease property proves the claim.

[3] A bi-tile T_i is a deterministic two tile complex T_{iA}, T_{iB}.

Next, we give an alternate binary encoding [33] for all non-zero natural numbers using $\{1, 2\}$ instead of the standard $\{0, 1\}$ encoding. This encoding will allow us to exploit the geometric decay property to build succinct constructions. The encoding of any non-zero natural number N is the N^{th} string in the lexicographic ordering of strings in $\{\mathbf{1}, \mathbf{2}\}^+$. An equivalent characterization s given below.

Lemma 4. *$\{\mathbf{1}, \mathbf{2}\}$-Binary Encoding: For all non-zero natural numbers N, $\exists b_i \in \{1, 2\} : N = \sum_{i=0}^{n-1} b_i 2^i$ where $n \leq \lceil \log N \rceil$. Every N has a unique $\{\mathbf{1}, \mathbf{2}\}$-binary encoding.*

Now we show how to encode any N using $\Theta(\log N)$ tile types using the above two Lemmas. Fig.3b is an example illustrating the construction for $N = 91$. For any given N, let N'' be the greatest even number less than or equal to N. For $N' = \frac{N''}{2}$, let $B(N') = b_{n-1}b_{n-2} \ldots b_0$ be its $\{\mathbf{1}, \mathbf{2}\}$-binary encoding of size n. For each bit b_i with $i \in \{0, 1, \ldots, n-2\}$, our construction has a tile complex T_{i+1} of size $2b_i$ tiles that occurs X_{i+1} times with $E[X_{i+1}] = 2^i$. For the bit b_{n-1}, the tile complex T'_n of size $2b_{n-1} - 1$ tiles occurs X'_n times with $E[X'_n] = 2^{n-1}$. Each time T'_n is deterministically preceded by either the seed or one of the bridge tiles. Each such complex is called T_n. Thus T_n is of size of $2b_{n-1}$ tiles and occurs X_n times with $E[X_n] = 2^{n-1}$. For odd N, we deterministically prefix a single tile to the West of the seed tile.

Theorem 2. *The above construction has an expected length $E[L] = N$ tiles and uses $\Theta(\log N)$ tile types.*

Proof. The length of the assembly L is given by $L = X_1 + X_2 + \cdots + X_n + (N \mod 2)$ and hence by linearity of expectation, $E[L] = 2(\sum_{i=0}^{n-1} b_i 2^i) + (N \mod 2) = N$. The number of tile types is $\Theta(n) = \Theta(\log N)$.

4.3 Lower Bounds on the Tile Complexity of Linear Assemblies of Expected Length N

In this section we prove that for all N the cardinality of any tile multiset that forms linear assemblies of expected length N in standard PTAM systems is $\Omega(\log N)$. The techniques that we use for deriving these tile complexity lower bounds are notable as they differ from traditional information theoretic methods used for lower bounds on tile complexity and furthermore our low bound results hold for each N, rather than for almost all N.

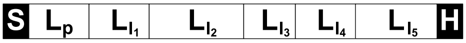


Fig. 4. $\mathbb{T}$ split into prefix and intermediates

Theorem 3. *For any N, the cardinality of any tile multiset that forms linear assemblies of expected length N in standard PTAM systems is $\Omega(\log N)$.*

Proof. We will show that any standard linear PTAM system with tile multiset cardinality n has expected length of assembly at most $O(2^n)$. This implies our result via the contrapositive. Recall that multiplicity of tiles in the multiset count distinctly towards tile complexity. Any standard PTAM linear tiling multiset with cardinality n that produces linear assemblies of greatest (finite) expected length is called *n-optimal.* Optimal linear tiling multisets must contain exactly one capping tile. If one had multiple capping tiles, say $term_1, \ldots, term_k$, replacing the East pads of $term_1, \ldots, term_{k-1}$ with the West pad of $term_k$ gives a modified tile multiset of same cardinality, which is still standard, and has a higher finite expected length, which is a contradiction. Define Ψ_n to be the expected length of the assembly produced by an n-optimal linear tiling multiset. We will prove $\Psi_n = O(2^n)$ by a recursive argument on n.

Let $\mathbb{T} = \langle T, \{s\}, g\rangle$ be any n-optimal linear tiling multiset. Let L be the random variable equal to the length of the linear assembly produced by $\mathbb{T}$ and so $E[L] = \Psi_n$. A *run* of a PTAM linear tiling system is a finite sequence of attachment of tile types resulting in a terminal assembly. A run might be alternatively thought of as a finite sequence of pad types where the number of pads in a run is one more than the number of tiles. For any run of $\mathbb{T}$, consider the pad type λ appearing on the West side of the capping tile. Let $\Lambda \subset T$ be the multiset of k_1 ($0 < k_1 < n - 1$[4]) tiles with λ as their West pad, not including the capping tile. Pad type λ might occur at many positions in this run. Define the *prefix* of the run as the subsequence from the West pad of the seed tile to the first occurrence of λ. Consider the subsequences that start and end in λ with no occurrence of λ within. Such a subsequence, excluding the first λ, is called an *intermediate* (See Fig.4). Define the following random variables: L_P equal to the length of the prefix, L_{I_i} equal to the length of the i^{th} intermediate subsequence and r equal to number of intermediates. The L_{I_i} are independent identical random variables and let L_I be a representative random variable with the same distribution. Length of the assembly equals the sum of the lengths of the prefix and the intermediates. Thus, $L = L_P + \sum_{i=1}^{r}(L_{I_i})$. For every i, the random variables r and L_{I_i} are independent because of the memoryless property of linear tiling systems. Thus, by linearity of expectation we get, $\Psi_n = E[L] = E[L_P] + E[\sum_{i=1}^{r}(L_{I_i})] = E[L_P] + E[r]E[L_I]$

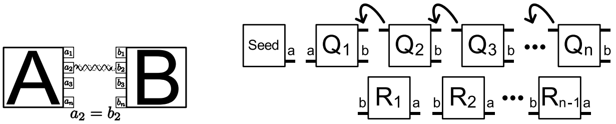
Since $\mathbb{T}$ is standard, each of the tiles in Λ and the capping tile can attach with equal probability $\frac{1}{k_1+1}$ to any tile with λ as its East pad. Thus, r is a geometric random variable, with parameter $\frac{1}{k_1+1}$, counting the number of times the capping tile fails to attach. Thus $E[r] = k_1$. We will show that $E[L_P]$ and $E[L_I]$ are at most Ψ_{n-k_1} by simulating the assemblies that produce these subsequences via linear tiling multisets of cardinality at most $n - k_1$. The prefix is simulated by the linear tiling system $\mathbb{T}_P$ obtained from $\mathbb{T}$ in the following manner. Drop the

[4] Note that Λ cannot be empty for an optimal linear tiling system. Suppose it were: let $\Lambda' = \{t_1, \ldots, t_k\}$, be the set of tile types with λ as their East pad. Replacing the East pads of $t_1, \ldots, t_{k-1}$ with the West pad of t_k gives a modified tile multiset of same cardinality, which is still standard, and has a higher finite expected length, which is a contradiction. The same arguments hold as we recurse. The seed s and the capping tile are never part of Λ.

tiles in Λ from $\mathbb{T}$. Observe that there is a run of $\mathbb{T}_P$ for every possible prefix and vice-versa, with the same probabilities of formation. Thus, the expected length of assembly produced by $\mathbb{T}_P$ is equal to $E[L_P]$. Also, the cardinality of tile multiset for $\mathbb{T}_P$ is $n - k_1$ and hence $E[L_P] \leq \Psi_{n-k_1}$ by definition. The intermediate sub-assemblies are simulated by a family of k_1 different tile systems. Each tile system has a tile multiset of cardinality $n - k_1$ obtained by $(i.)$ dropping the tiles in Λ from $\mathbb{T}$ and $(ii.)$ replacing the seed tile by some $t \in \Lambda$ and making $\text{pad}_{\text{West}}(t) = \phi$. Each intermediate sub-assembly is simulated by some tile system from this family. Thus $E[L_I] \leq \Psi_{n-k_1} - 1$. Thus, $\Psi_n = E[L_P] + E[r]E[L_I] \leq (k_1+1)\Psi_{n-k_1}$. In the next level of recursion, we drop $k_2 > 0$ tiles to get $\Psi_n \leq (k_1+1)\Psi_{n-k_1} \leq (k_1+1)(k_2+1)\Psi_{n-k_1-k_2}$. In general, we drop k_i tiles in the i^{th} level of recursion to get $\Psi_n \leq \prod_{j=1}^{i}(k_j+1)\Psi_{n-\sum_{j=1}^{i} k_j}$. The base case is $\Psi_2 = 2$ since the best one can do with a single seed and capping tile is assembly of length 2. Also, let there be z levels of recursion. Thus $\Psi_n \leq \prod_{i=1}^{z}(k_i+1)$ with $\sum_{i=1}^{z} k_i = n-2$. The product $\prod_{i=1}^{z}(k_i+1)$ constrained by $\sum_{i=1}^{z} k_i = n-2$ has a maximum value of 2^{n-2}. Hence $\Psi_n \leq O(2^n)$.

5 κ-Pad Systems for Linear Assembly

In this section we will extend PTAM by modifying each tile to accommodate multiple pads on each side. Tiles bind when one pair of adjacent pads match (see Fig.5a). To ensure that tiles align fully and are not offset, each pad on a side of a tile are drawn from different sets of pad types. Using such multi-padded tiles, we will show it is possible to reduce the number of tile types to get linear assemblies of expected length N.



(a) (b)

Fig. 5. κ-pad Systems. (a) κ-pad tiles A and B. (b) Pad binding diagram for linear tiling system using $\Theta_{i.o}(\frac{\log N}{\log \log N})$ 2-pad tile types. Small labeled rectangles on the sides of the tiles indicate various types of pads. Arrows indicate possible attachment. Absent pads are ϕ.

5.1 Definitions

A *κ-pad tile* t over the cartesian product $\Sigma = \Sigma_1 \times \Sigma_2 \times \cdots \times \Sigma_\kappa$ is a unit square whose two opposite sides each have a κ tuple of pads from Σ. Thus, tile $t \in T$ is an ordered pair[5] (W_t, E_t) where W_t and E_t are row vectors of size κ, where the i^{th} component of each vector is from the set Σ_i. Thus, the East

[5] Again, for two dimensional assemblies, tiles have pads on all four sides and the model can be extended to include a temperature parameter τ for co-operative binding interactions with multiple tiles.

and West sides of each tile has κ pads. $\Sigma_1, \ldots, \Sigma_\kappa$ are finite, mutually disjoint set of distinct pad types. A *κ-pad* linear tiling system $\mathbb{T}$ is given by the tuple $\langle T, S, g \rangle$ where T is the finite multiset of κ-pad tile types, $S \subset T$ is the set of seed tiles and g is the binary pad strength function. Definitions from Section 3 hold with appropriate modifications to incorporate multiple pads on sides of each tile. For each tile t, we define $\text{pad}_{\text{East}}(t, i) = (E_t)_i$ and $\text{pad}_{\text{West}}(t, i) = (W_t)_i$ where $(E_t)_i$ and $(W_t)_i$ denote the i^{th} component of the respective pad vectors. For $D \in \mathfrak{D}$ we say the tiles at x and $D(x)$ *attach* if there exists an i such that $g(\text{pad}_D(A(x), i), \text{pad}_{D^{-1}}(A(D(x)), i)) = 1$. (See Fig.5a).

With these modifications, *diagonal*, *uni-seeded* and *haltable* linear tiling systems and self-assembly of κ-pad tiles are defined as in Sections 2 and 3. In particular, *probabilities of attachment* of tiles is given by the same formula as in Section 3 and Lemma 1 holds for κ-pad systems. We restrict ourselves to studying diagonal, uni-seeded and haltable κ-pad linear tiling systems. Note that for assemblies in Section 5.3, adjacent tiles that bind have exactly one match among corresponding pads.

5.2 Implementing κ-Pad Systems Using DNA Self-assembly

κ-pad tiles can be feasibly realized using carefully designed self-assembled DNA motifs. Indeed, the DX motif [10], one of the early demonstrations of DNA motifs that self-assemble into two dimensional lattices, can serve as a 2-pad tile. Other similar motifs that also self-assemble into two dimensional lattices, like the TX [26], can serve as multipad systems. These motifs can be easily modified to self-assemble in one dimension, as a linear structure. On a much larger scale, Rothemund's origami technique [1] can be used to manufacture tiles with hundreds of pads. A drawback of such a system would be that the connection between adjacent tiles will be quite flexible, making a linear assembly behave more as a chain rather than a rigid ruler.

5.3 Linear Assemblies of Expected Length N Using $\Theta_{i.o}(\frac{\log N}{\log \log N})$ 2-Pad Tile Types

In this section we present an equimolar, standard κ-pad linear tiling system with $\kappa = 2$, i.e a 2-pad system, that achieves for any given $N' \in \mathbb{N}$, a linear assembly of expected length $N > N'$ using $\Theta(\frac{\log N}{\log \log N})$ 2-pad tile types, i.e., arbitrary long fixed length assemblies of expected length N using $\Theta(\frac{\log N}{\log \log N})$ 2-pad tile types. Fig.5b illustrates the tile set used in our construction. $Q_2, Q_3 \ldots Q_n$ are bi-tiles with deterministic internal pads and so for simplicity we will treat them as a single tile of length two. R is a tile type with multiplicity $n-1$, drawn as $R_1, \ldots, R_{n-1}$ in Fig.5b. Q_{i+1} can attach to Q_i's East side via the upper pad. For $j \in \{1, 2, \ldots, n-1\}$, $R_1, R_2, \ldots, R_{n-1}$ can attach to Q_j's East side via the lower pad and Q_1 attaches deterministically to R_j's East side via the lower pad. Q_1 attaches deterministically to the seed's East side while Q_n is the capping tile. The assembly halts iff the consecutive sequence $Q_1, Q_2, \ldots, Q_n$ occurs. At each stage, the assembly can restart by the attachment of Q_1 via any of the $n-1$ bridge tiles R_j. The number of tile types is $2n = \Theta(n)$.

Theorem 4. *Let X be the random variable that equals the length of the tile system illustrated in Fig.5b. Then $E[X] = N = \Theta(n^n)$ using $\Theta(\frac{\log N}{\log \log N}) = \Theta(n)$ 2-pad tile types.*

Proof. We can think of the process as a series of Bernoulli trials, say biased coin tosses. A *Head* corresponds to attachment of some Q_i $(i \neq 1)$ and a *Tail* to some R_j, Q_1 complex. The probability of a *Head* is $\frac{1}{n}$. The assembly halts iff $n-1$ successive heads occur. Each toss adds exactly 2 tiles to the assembly and the seed and Q_1 appear once before the first toss. So, from [32], the expected length of the assembly is given by $E[X] = N = \Theta(n^n)$. The number of tile types used is $\Theta(n) = \Theta(\frac{\log N}{\log \log N})$.

5.4 Lower Bounds for κ-Pad Systems

In this section we prove for each N that the cardinality of κ-pad tile multiset required to form linear assemblies of expected length N in standard PTAM systems is $\Omega(\frac{\log N}{\log \log N})$. Prior self-assembly lower bounds on numbers of tiles for assembly used information theoretic methods, whereas this proof is via a reduction to a problem in linear algebra, and furthermore holds for each N, rather than for almost all N.

Theorem 5. *For each N, the cardinality of the smallest κ-pad tile multiset required to form linear assemblies of expected length N in standard PTAM systems is $\Omega(\frac{\log N}{\log \log N})$.*

Proof. As in the Theorem 3, we will show that any κ-pad standard linear PTAM system with tile multiset of cardinality n has expected length of assembly at most $O(n^{2n})$ and this implies our result via the contrapositive. The proof uses a reduction to determining the expected time to first arrival at a vertex in a random walk over a graph, which is further reduced to a problem in linear algebra, namely determining magnitude bounds on the solution of a linear system, which is bounded by the magnitude of a ratio of two determinants.

Any n-optimal κ-pad system $\mathbb{T} = \langle T, \{s\}, g \rangle$ has exactly one seed and one capping tile, by an argument similar to the one in Section 4.3. Let L be the random variable equal to the length of linear assembly produced by $\mathbb{T}$. Consider the directed weighted graph $G = (V, E, w)$ constructed from $\mathbb{T}$ as follows: *i.* V is in one-to-one correspondence with T where vertices in V have distinct labels for repeated tile types in T, *ii.* directed edge $(u, v) \in E$ iff the East face of tile corresponding to u and West face of tile corresponding to v can attach and *iii.* for each $(u, v) \in E$, edge weights indicating transition probability are given by $w(u, v) = (\text{outdegree(u)})^{-1}$. Note that the sum of edge weights of all edges leaving a node is 1 and all edges leaving a vertex have equal transition probability. G has a *start* vertex corresponding to the seed s and a *destination* vertex corresponding to the capping tile. Self-assembly is a random walk on G

from the start to the destination, where paths in G from the start correspond to produced configurations in $\mathbb{T}$. Let *expected time to destination*, $\delta(u)$, be the expected length of the random walk from u to destination for some $u \in V$. The expected length of the assembly, $E[L] = \delta(start)$ and $\delta(destination) = 0$.

For any $u \in V$ (other than the destination), with edges to the k vertices $\{v_1, \ldots, v_k\}$, $\delta(u) = 1 + \sum_{i=1}^{k} \frac{1}{k}\delta(v_i)$. Writing such equations for each vertex in V, with $\delta(destination) = 0$ gives a system of n linear equations in n variables, say $\mathbf{A}\delta = \mathbf{b}$ where $\mathbf{A}$ is an $n \times n$ matrix of transition probabilities with values from the set $\{0, \frac{1}{n}, \frac{1}{n-1}, \ldots, 1\}$, $\mathbf{b} = [\mathbf{1}\ \mathbf{1}\ \ldots\ \mathbf{1}\ \mathbf{0}]^{\mathrm{T}}$ is a vector of size n and δ is the vector of expected times to destination. $\mathbf{A}$ is non-singular and therefore the system has a unique solution[6]. Using Cramer's rule $\delta(s) = \frac{|\mathbf{A_b}|}{|\mathbf{A}|}$ where $\mathbf{A_b}$ is the appropriate column of $\mathbf{A}$ substituted by $\mathbf{b}$. We upper and lower bound the two determinants using Leibniz's formula, $|C| = \sum_{\pi \in S_n} sgn(\pi) \prod_{i=1}^{n} C_{i,\pi(i)}$ where the sum is computed over all $n!$ permutations of S_n, where S_n is the permutations of the set $\{1, 2, \ldots, n\}$ and $sgn(\sigma)$ denotes the signature of the permutation σ: $+1$ if σ is an even permutation and -1 if it is odd. Note that the maximum value of the product $\prod_{i=1}^{n} C_{i,\pi(i)}$ is 1 since the values in each of the determinants are from the set $\{0, \frac{1}{n}, \frac{1}{n-1}, \ldots, 1\}$. Thus $|\mathbf{A_b}| \leq n!$ and similarly $|\mathbf{A}| \geq (1/n)^n$. Hence, $\delta(s) \leq O(n^{2n})$. Thus the expected length of an assembly of any κ-pad standard linear PTAM system with tile multiset of cardinality n is at most $O(n^{2n})$ which implies a lower bound of $\Omega(\frac{\log N}{\log \log N})$.

6 Improving Tail Bounds of Distribution of Lengths of Assembly

Linear tile systems that do not give assemblies with exponential tail bounds on length can be modified by concatenating k independent, distinct versions of the tile system into a new tile system with tail bounds that drop exponentially with k. Both the central limit theorem and Chernoff bounds are used for bounding the tail of this new distribution.

Given a tile multiset T (with single or κ-pads on each side of each tile) for a linear assembly, let $\hat{L}$ be the random variable equal to the length of the assembly with mean $\lfloor \frac{N}{k} \rfloor$ and variance $\frac{\sigma^2}{k}$, and let $f(\lfloor \frac{N}{k} \rfloor)$ be the cardinality of T. Consider k distinct versions of T, say $T_1, T_2, \ldots, T_k$, each mutually disjoint. We deterministically concatenate the assemblies produced by these tile multisets by introducing pads that allow the East side of each capping tile of T_i to attach to the West side of the seed tile of T_{i+1} for $i = 1, 2, \ldots, n-1$. We then add $N - k\lfloor \frac{N}{k} \rfloor \leq k$ distinct tiles that deterministically extend the assembly beyond the capping tile of T_k. Let L the random variable equal to the length of the assembly produced by this construction. This new multiset, T_{sh} of cardinality $f_{\mathrm{sh}}(N) \leq kf(\lfloor \frac{N}{k} \rfloor) + k$ gives linear assemblies of expected length $E[L] = N$ and variance σ^2. $k \in \{1, \ldots, N\}$ determines how sharp the overall probability distribution is.

[6] The solution is unique as the expected number of transitions from any vertex to the capping vertex is well defined.

The central limit theorem gives: $\forall \delta \geq 0 : P(|L - N| \leq \delta\sigma) \rightarrow \Phi(\delta)$ as $k \rightarrow \infty$, where Φ and ψ are the probability density function and cumulative distribution function respectively of the standard normal distribution. Thus, $P(|L - N| \geq \delta\sigma) \rightarrow 2(1 - \Phi(\delta)) \leq 2\psi(\delta)/\delta \leq \sqrt{2/\pi}(e^{-\delta^2/2}/\delta)$ as $k \rightarrow \infty$. Thus, we achieve an exponentially decaying tail bound with a linear multiplicative increase in tile complexity for large k. Since T_{sh} is the concatenation of independent assemblies T_i, Chernoff bounds for sums of independent random variables gives $\forall \delta, t > 0 : P(L > (1 + \delta)N) \leq (M(t)/e^{(1+\delta)\lfloor \frac{N}{k} \rfloor t})^k$ and $\forall \delta > 0, t < 0 : P(L < (1 - \delta)N) \leq (M(t)/e^{(1-\delta)\lfloor \frac{N}{k} \rfloor t})^k$ where $M(t)$ is the moment generating function of the random variable $\hat{L}$. If $M(t)/e^{(1+\delta)\lfloor \frac{N}{k} \rfloor t} < 1$ for some $t > 0$ and $M(t)/e^{(1-\delta)\lfloor \frac{N}{k} \rfloor t} < 1$ for some $t < 0$, we get tail bounds dropping exponentially with k.

Acknowledgements

We wish to thank our anonymous reviewers for helpful comments, especially for pointing out the error resilience of temperature 1 assemblies. We also thank Josh Letchford and Thom LaBean for helpful discussions and Sudheer Sahu, Manoj Gopalkrishnan, Jeff Phillips, Shashidhara Ganjugunte and Urmi Majumder for their comments on an earlier draft of this paper. This work was supported by NSF EMT Grant CCF-0829797, CCF-0829798 and AFSOR Contract FA9550-08-1-0188.

References

1. Rothemund, P.: Folding DNA to Create Nanoscale Shapes and Patterns. Nature 440, 297–302 (2006)
2. Yan, H., Yin, P., Park, S.H., Li, H., Feng, L., Guan, X., Liu, D., Reif, J., LaBean, T.: Self-assembled DNA Structures for Nanoconstruction. American Institute of Physics Conference Series, vol. 725, pp. 43–52 (2004)
3. Zhang, D.Y., Turberfield, A., Yurke, B., Winfree, E.: Engineering Entropy-Driven Reactions and Networks Catalyzed by DNA. Science 318, 1121–1125 (2007)
4. Wang, H.: Proving Theorems by Pattern Recognition II (1961)
5. Berger, R.: The Undecidability of the Domino Problem, vol. 66, pp. 1–72 (1966)
6. Papadimitriou, C.: Computational Complexity. Addison-Wesley, Reading (1993)
7. Robinson, R.: Undecidability and Nonperiodicity for Tilings of the Plane. Inventiones Mathematicae 12, 177–209 (1971)
8. Garey, M., Johnson, D.: Computers and Intractability: A Guide to the Theory of NP-Completeness. W.H. Freeman, New York (1981)
9. Lewis, H., Papadimitriou, C.: Elements of the Theory of Computation. Prentice-Hall, Englewood Cliffs (1981)
10. Winfree, E., Liu, F., Wenzler, L., Seeman, N.: Design and Self-Assembly of Two-Dimensional DNA Crystals. Nature 394, 539–544 (1999)
11. Adleman, L.: Towards a mathematical theory of self-assembly. Technical report, University of Southern California (2000)
12. Winfree, E.: DNA Computing by Self-Assembly. In: NAE's The Bridge, vol. 33, pp. 31–38 (2003)
13. Rothemund, P., Winfree, E.: The Program-Size Complexity of Self-Assembled Squares. In: STOC, pp. 459–468 (2000)

14. Adleman, L., Cheng, Q., Goel, A., Huang, M.D.: Running Time and Program Size for Self-Assembled Squares. In: STOC, pp. 740–748 (2001)
15. Barish, R., Rothemund, P., Winfree, E.: Two Computational Primitives for Algorithmic Self-Assembly: Copying and Counting. Nano Letters 5(12), 2586–2592 (2005)
16. Yan, H., Feng, L., LaBean, T., Reif, J.: Parallel Molecular Computation of Pair-Wise XOR using DNA String Tile. Journal of the American Chemical Society (125) (2003)
17. Winfree, E.: Simulations of Computing by Self-Assembly. Technical report, Caltech CS Tech Report (1998)
18. Schulman, R., Lee, S., Papadakis, N., Winfree, E.: One Dimensional Boundaries for DNA Tile Self-Assembly. In: Chen, J., Reif, J.H. (eds.) DNA 2003. LNCS, vol. 2943, pp. 108–126. Springer, Heidelberg (2004)
19. Park, S.H., Yin, P., Liu, Y., Reif, J., LaBean, T., Yan, H.: Programmable DNA Self-assemblies for Nanoscale Organization of Ligands and Proteins. Nano Letters 5, 729–733 (2005)
20. Aggarwal, G., Goldwasser, M., Kao, M.Y., Schweller, R.: Complexities for Generalized Models of Self-Assembly. In: SODA, pp. 880–889 (2004)
21. Kao, M.Y., Schweller, R.: Randomized Self-Assembly for Approximate Shapes. In: Aceto, L., Damgård, I., Goldberg, L.A., Halldórsson, M.M., Ingólfsdóttir, A., Walukiewicz, I. (eds.) ICALP 2008, Part I. LNCS, vol. 5125, pp. 370–384. Springer, Heidelberg (2008)
22. Lagoudakis, M., LaBean, T.: 2D DNA Self-Assembly for Satisfiability. In: DIMACS Workshop on DNA Based Computers (1999)
23. Becker, F., Rapaport, I., Rémila, É.: Self-assemblying classes of shapes with a minimum number of tiles, and in optimal time. In: Arun-Kumar, S., Garg, N. (eds.) FSTTCS 2006. LNCS, vol. 4337, pp. 45–56. Springer, Heidelberg (2006)
24. Reif, J.: Local parallel biomolecular computation. In: NA-Based Computers, III, pp. 217–254. American Mathematical Society, Providence, RI (1997)
25. Demaine, E.D., Demaine, M.L., Fekete, S.P., Ishaque, M., Rafalin, E., Schweller, R.T., Souvaine, D.L.: Staged Self-assembly: Nanomanufacture of Arbitrary Shapes with $O(1)$ Glues. In: Garzon, M.H., Yan, H. (eds.) DNA 2007. LNCS, vol. 4848, pp. 1–14. Springer, Heidelberg (2008)
26. LaBean, T., Yan, H., Kopatsch, J., Liu, F., Winfree, E., Reif, J., Seeman, N.: Construction, Analysis, Ligation, and Self-Assembly of DNA Triple Crossover Complexes. Journal of the American Chemical Society 122(9), 1848–1860 (2000)
27. Mao, C., LaBean, T., Reif, J., Seeman, N.: Logical Computation Using Algorithmic Self-Assembly of DNA Triple-Crossover Molecules. Nature 407, 493–496 (2000)
28. Gillespie, D.: Exact Stochastic Simulation of Coupled Chemical Reactions. The Journal of Physical Chemistry 81, 2340–2361 (1977)
29. Soloveichik, D., Winfree, E.: Complexity of Self-Assembled Shapes. SIAM Journal of Computing 36(6), 1544–1569 (2007)
30. Winfree, E., Bekbolatov, R.: Proofreading Tile Sets: Error Correction for Algorithmic Self-Assembly. In: Chen, J., Reif, J.H. (eds.) DNA 2003. LNCS, vol. 2943, pp. 126–144. Springer, Heidelberg (2004)
31. Chen, H.-L., Goel, A.: Error free self-assembly using error prone tiles. In: Ferretti, C., Mauri, G., Zandron, C. (eds.) DNA 2004. LNCS, vol. 3384, pp. 62–75. Springer, Heidelberg (2005)
32. Gordan, H.: Discrete Probability. Springer, Heidelberg (1997)
33. Li, M., Vitanyi, P.: An Introduction to Kolmogorov Complexity and Its Applications, 2nd edn. Springer, Heidelberg (1997)