TILE COMPLEXITY OF LINEAR ASSEMBLIES[§]

HARISH CHANDRAN[*], NIKHIL GOPALKRISHNAN[†], AND JOHN REIF[‡]

Abstract. Self-assembly is fundamental to both biological processes and nanoscience. Key features of self-assembly are its probabilistic nature and local programmability. These features can be leveraged to design better self-assembled systems. 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. A particular challenge in DNA nanoscience is to form linear assemblies or *rulers* of a specified length using the smallest possible tile set, where any tile type may appear more than once in the assembly. The tile complexity of a linear assembly is the cardinality of the tile set that produces it. These rulers can then be used as components for construction of other complex structures. While square assemblies have been extensively studied, many questions remain about fixed length linear assemblies, which are more basic constructs yet fundamental building blocks for molecular architectures. In this work, we extend TAM to take advantage of inherent probabilistic behavior in physically realized self-assembled systems by introducing randomization. We describe a natural extension to TAM called the Probabilistic Tile Assembly Model (PTAM). A restriction of the model, which we call the standard PTAM is considered in this report. Prior work in DNA self-assembly strongly suggests that standard PTAM can be realized in the laboratory. In TAM, a deterministic linear assembly of length N requires a tile set of cardinality at least N. In contrast, we show various non-trivial probabilistic constructions for forming linear assemblies in PTAM with tile sets of sub-linear cardinality, using techniques that differ considerably from existing assembly techniques. In particular, for any given N, we demonstrate linear assemblies of expected length N with a tile set of cardinality $\Theta(\log N)$ using one pad per side of each tile. We prove 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. We further demonstrate how linear assemblies can be modified to produce assemblies with sharp tail bounds on distribution of lengths by concatenating various assemblies together. In particular, we show that for infinitely many N we can get linear assemblies with exponentially dropping tail distributions using $O(\log^3 N)$ tile types. 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. This gives linear assemblies of expected length N with a 2-pad (two pads per side of each tile) tile set of cardinality $\Theta\left(\frac{\log N}{\log\log N}\right)$ for infinitely many N. We show that we cannot get smaller tile complexity by proving a lower bound of $\Omega\left(\frac{\log N}{\log\log N}\right)$ 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 positive integer κ. The techniques that we use for deriving these tile complexity lower bounds are notable as they differ from traditional Kolmogorov complexity based information theoretic methods used for lower bounds on tile complexity. Also, Kolmogorov complexity based lower bounds do not preclude the possibility of achieving assemblies of very small tile multiset cardinality for infinitely many N. In contrast, our lower bounds are stronger as they hold for every N, rather than for almost all N. All our probabilistic constructions are free from co-operative tile binding errors. 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.

Keywords: Tile Assembly Model, Tile complexity, Linear assemblies, Wang tilings, Self-assembly, DNA tiles.

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 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 techniques have produced increasingly complex structures (see Park et al. (2005); Rothemund (2006); Douglas et al. (2009); Dietz et al. (2009); Zheng et al. (2009); Andersen et al. (2009) for a few illustrative examples) and dynamical systems (see Dirks

[§] A PRELIMINARY VERSION OF THIS WORK HAS APPEARED IN 36TH INTERNATIONAL COLLOQUIUM ON AUTOMATA, LANGUAGES AND PROGRAMMING (ICALP 2009).

[*] Department of Computer Science, Duke University, Durham, NC. email: harish@cs.duke.edu

[†] Department of Computer Science, Duke University, Durham, NC. email: nikhil@cs.duke.edu

[‡] Department of Computer Science, Duke University, Durham, NC and Adjunct, Faculty of Computing and Information Technology (FCIT), King Abdulaziz University (KAU), Jeddah, Saudi Arabia. email: reif@cs.duke.edu

and Pierce (2004); Zhang et al. (2007); Yin et al. (2008) for some examples). The roots of attempts to model and study self-assembly begin with the study of tilings.

A *Wang tile* (Wang (1961)), is an oriented unit square with a pad associated with each side. Any two tiles with the same pads 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 (1966) proved the undecidability of the tiling problem by reducing the halting problem to it. Robinson (1971) gave an alternative proof involving a simulation of any single tape deterministic Turing Machine by some set of Wang tiles. Garey and Johnson (1981) and Lewis and Papadimitriou (1981) 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 (1995) extended Wang tilings to the *Tile Assembly Model* (TAM) with a view to model self-assembly processes, laying a theoretical foundation (see Winfree (1998a); Adleman (2000); Rothemund and Winfree (2000)) for a form of DNA based computation, particularly, molecular computation via assembly of DNA lattices with tiles in the form of DNA motifs.

The *tile complexity*, defined first by Rothemund and Winfree (2000), 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 (see Rothemund and Winfree (2000); Adleman et al. (2001a); Kao and Schweller (2008); Doty (2009)), rectangles (see Aggarwal et al. (2004)) and computations like counting (see Barish et al. (2005)), Sierpinski triangles (Rothemund et al. (2004)) etc. exist in this model. Lower bounds on tile set complexity have also been shown for various shapes (see Rothemund and Winfree (2000); Aggarwal et al. (2004); Doty et al. (2011)).

Stochastic processes play a major role in self-assembly and have been investigated theoretically by Winfree (1998b) and Adleman (2000) and in the laboratory by Schulman and Winfree (2007). However, constructions in TAM are typically deterministic in the sense that they produce exactly one terminal assembly given a tile set (see Bryans et al. (2011) for non-deterministic constructions in TAM). 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 as a component for construction of other complex structures. One can use these structures as nanoscale beams and struts (See Fig.1.1).

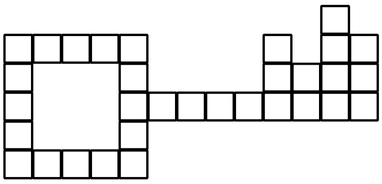

Fig. 1.1: *Possible nanostructures using rulers as substructures.*

Linear assemblies can also serve as boundaries as demonstrated by Schulman et al. (2004) and as nucleation sites for more complex nanostructures. Note that due to the inherently flexible nature of linear nanostructures, most complex nanostructures will generally tolerate small deviations from the intended lengths of

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

these substructures. 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, the number of tile types to form an $N \times N$ square is only $\Theta\left(\frac{\log N}{\log \log N}\right)$ as proved by Adleman et al. (2001a), 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 (see Rothemund and Winfree (2000)) while the one for linear assemblies is not. This begs the question: why are we not able to reach the information theoretic limit of $\Theta\left(\frac{\log N}{\log \log N}\right)$ 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 (see Rothemund and Winfree (2000); Adleman et al. (2001a); Kao and Schweller (2008); Doty (2010)), 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 the 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.

1.3. Related Work in Self-Assembly using Probabilistic and Randomized Models. Non-deterministic tilings were studied by Lagoudakis and LaBean (1998) for implementing an algorithm for SAT. Becker et al. (2006) describe probabilistic tile systems that yield squares, rectangles and diamonds in expectation using $O(1)$ tile types. This work was extended by Kao and Schweller (2008) to yield arbitrarily close approximations to squares with arbitrarily high probability using $O(1)$ tile types. Doty (2010) improved the techniques developed by Kao and Schweller (2008) to get squares with arbitrarily high probability using $O(1)$ tile types. These works require 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, the size of the tile set producing an assembly is not a true indicator of its descriptional complexity. In PTAM, the set of tiles is a multi-set that implicity defines relative concentrations and thus imposes a charge for specifying relative concentrations. Therefore, the size of the tile set producing an assembly is a true indicator of its descriptional complexity.

Demaine et al. (2008) discuss *staged self-assembly* to get various shapes using $O(1)$ pad types. Aggarwal et al. (2005) 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.

The Kinetic Tile Assembly Model (kTAM) proposed by Winfree (1998b) models kinetics and thermodynamics of DNA hybridization reactions. Schulman et al. (2004) used the DNA based DX tiles, originally designed by Winfree et al. (1998), to create one dimensional boundaries within the nanoscale. Adleman (2000) 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 article are original and the constructions presented are novel.

1.4. Main Results. We describe a natural extension to TAM in Section 3 to allow stochastic, non-deterministic 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 article. Prior work in DNA self-assembly strongly suggests that standard PTAM constructs can be realized in the laboratory. We show various non-trivial probabilistic constructions in PTAM for forming linear assemblies with small tile sets in Section 4, using techniques that differ considerably from existing assembly techniques. In Section 4.2, 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 5 we demonstrate how linear assemblies can be modified to produce assemblies with sharp tail bounds on distribution of lengths by concatenating various assemblies together. In particular, in Section 5.2.3, we show that for infinitely many N we can get linear assemblies with exponentially dropping tail distributions using $O(\log^3 N)$ tile types. We derive a 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 6. This lower bound, which holds for all N, is tight and stronger than the information theoretic lower bound of $\Omega\left(\frac{\log N}{\log\log N}\right)$ which holds only for almost all N. We also propose a simple extension to PTAM in Section 7 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 a 2-pad (two pads per side of each tile) tile set of cardinality $\Theta\left(\frac{\log N}{\log\log N}\right)$ tile types for infinitely many N as proved in Section 7.3. We show in Section 7.4 that we cannot achieve smaller tile complexity by proving a lower bound of $\Omega\left(\frac{\log N}{\log\log N}\right)$ for each N on the cardinality of the κ-pad (κ pads per side of each tile) tile multiset 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.

2. The Tile Assembly Model for Linear Assemblies. This section describes the Tile Assembly Model (TAM) by Winfree for special case linear (1D) assemblies (henceforth referred to as LTAM). For a complete and formal description of the general model see Rothemund and Winfree (2000). The next section extends the model by introducing stochasticity and non-determinism. This article considers only a 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. A function $g : \Sigma \times \Sigma \to \{0, 1\}$ 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 during an attachment step. Hence the temperature parameter used in TAM is redundant in linear assemblies where tiles have only one pad per side. Throughout this article 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 7 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 tiles, $S \subset T$ is the set of *seed* tiles and g is the binary pad strength function. A *configuration* of T is a function $A : \mathbb{Z} \to T$ 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, $A \to B$, 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 $\to$ 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 $\forall x \neq 0 : \text{start}_s(x) = \text{empty}$. A configuration B is *produced*

[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 article and hence omit them. Also, we allow for multiple pads on the sides of a tile in Section 7.

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.

DNA nanostructures can physically realize TAM as shown by Winfree et al. (1998) with the DX tile and Mao et al. (2000) 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. (2000) performed a laboratory demonstration of computation via tile assembly using TX tiles. Yan et al. (2003) performed parallel XOR computation in the test-tube using DX tiles. Other simple computations have also been demonstrated. However, larger and more complex computations are beset by errors and correction of these errors remains a challenge towards general computing using DNA tiles. Winfree and Bekbolatov (2003); Chen and Goel (2004); Reif et al. (2004); Chen et al. (2007) design basic error correction protocols. See Section 8 for a further discussion.

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 article, 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 attachment at certain positions in a configuration. Note that we do not allow pad mismatch errors (see Section 8 for details). 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 the North and South sides of tiles and including a temperature parameter τ as defined by Rothemund and Winfree (2000) 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, $S \subset T$ 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* $\mathcal{M} : T \to \mathbb{N}$ of a tile is the number of times it occurs in T. T contains the empty tile type with $\mathcal{M}(\textit{empty}) = 1$. Multiplicity models concentration. We assume a well-mixed reaction environment in which, at each step, some member of T is copied (chosen with replacement) with uniform probability. If the tile thus obtained can attach to the produced configuration, it does so, else we re-sample from T 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 (see Gillespie (1977)) 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 article. The function *type(t)*, $\text{type} : T \to \Sigma \times \Sigma$, returns the tile type for any $t \in T$.

Self-assembly of a linear tiling system $\mathbb{T}$ is defined by a relation between the set of positive probabilities and a pair of configurations A and B as: $A \to^p_{\mathbb{T}} B$ (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 $\to^p_{\mathbb{T}}$, denoted by $\overset{*}{\to}{}^{\hat{p}}_{\mathbb{T}}$ (read as 'derives'), is defined by the following transitive law: if $A \to^{p_1}_{\mathbb{T}} B$ and $B \to^{p_2}_{\mathbb{T}} C$ then $A \to^{p_1 p_2}_{\mathbb{T}} C$. A configuration B is *produced* with positive probability p if $\text{start}_s \overset{*}{\to}{}^{p}_{\mathbb{T}} B$. A configuration is *terminal* if it is produced from start_s and no other configuration can be derived from it with positive probability. $\text{Term}(\mathbb{T})$ 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 $\Gamma = \{k|A_k \to^{p_k}_{\mathbb{T}} B\}$. *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 terminology developed by Rothemund and Winfree (2000), 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 $A \overset{*}{\rightarrow}{}^{p}_{\mathbb{T}} B$ 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 below.

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

Proof. Consider the directed weighted graph G whose nodes are produced configurations. Designate the node corresponding to the start configuration start_s as the *start* node of G. An edge exists from a node A to node B with probability of transition p iff $A \rightarrow^{p}_{\mathbb{T}} B$. Note that G might have infinite number of nodes. Let the nodes of G with outdegree 0 be called *leaf* nodes. $Term(\mathbb{T})$ is in one-to-one correspondence with the leaf nodes of this tree. The probability of formation for any produced configuration can be read out from its corresponding node by summing the product of transition probabilities over all paths from the start to that node. We will show a conservation law for the probability of formation which will inductively imply that the sum of probabilities of formation of leaf nodes equals the probability of formation of the start node, which is $P(\text{start}_s) = 1$. Let us partition the set of nodes of G into *levels* corresponding to the breadth-first traversal of G from the start node. Level 0 contains only the start node. Level i contains all nodes that are i hops away from the start node. Note that each node at level i corresponds to a configuration having exactly $i + 1$ non-empty tiles. Since a configuration having $i + 1$ non-empty tiles can only be formed by attachment of a single tile to a configuration having i tiles, each edge of G is across consecutive levels. The conservation law is, sum of probabilities of formation of non-leaf nodes at level i equals sum of probabilities of formation of nodes at level $i + 1$, which follows directly from the recursive definition of the probability of formation and the above described special structure of G. Thus, if $\mathbb{T}$ is haltable, this conservation law guarantees that sum of probabilities of formation of leaf nodes is $P(\text{seed}_s) = 1$. Thus $\displaystyle\sum_{A \in Term(\mathbb{T})} P(A) = 1.$ □

A linear tiling system is called *east-growing* if the West pad of the seed tile is ϕ. A *simulation* of a probabilistic linear tile system $\mathbb{T}$ by a probabilistic linear 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 $c = (W_c, \phi)$ 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 report. 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.

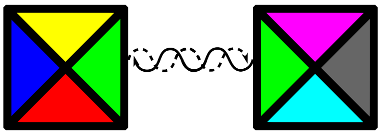

Fig. 3.1: *Diagonal tiles: Colors indicate pad type. Green pads are implemented using complementary DNA. Strands for other pads arenot shown.*

Diagonal tile systems were suggested by Rothemund and Winfree (2000). These systems are implementable using DNA tiles. Matching pads are implemented as perfect Watson-Crick complementary DNA sequences (see Fig.3.1). Non-diagonal tile systems are implementable using κ-pad systems with diagonal glue strength functions. 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 article. Achieving arbitrary concentration vectors is infeasible in laboratory implementations using molecules. In contrast, equimolar systems, or close approximations to them, are frequently achieved by chemists for various reactions. We demonstrate an 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 is defined as the number of different tile types in the smallest tile set that realizes the shape. 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. The tile complexity in TAM is closely related to the size of the smallest Turing machine describing the shape (see Soloveichik and Winfree (2007) for results connecting *scale-free* tile complexity and Kolmogorov complexity of that shape). However there exist modifications of TAM (see Aggarwal et al. (2004); Kao and Schweller (2006); Demaine et al. (2007); Becker et al. (2006)) where the number of tile types do 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. Thus, the (probabilistic) descriptional complexity of shapes corresponds to the cardinality of the tile multiset which we call tile complexity. Note that the 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 many times. 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. We focus on the expected lengths of linear assemblies in this section. In Section 5 we discuss methods to achieve a sharp distribution around the expectation. All of our constructions for linear assemblies of expected length $N \in \mathbb{N}$ in this section 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, except when otherwise indicated. 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 (see Winfree and Bekbolatov (2003); Chen and Goel (2004)). 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 equals 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 linear assemblies of expected length N 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.

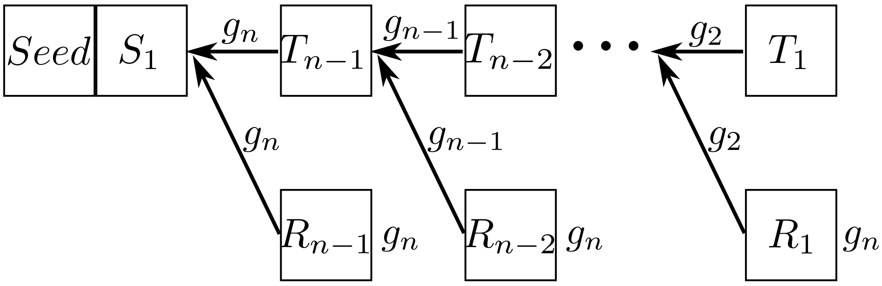

Fig. 4.1: *Tile binding diagram for powers of two construction. The labels on arrows indicate pad type. The arrows for the label g_n on tiles R_i are not drawn for the sake of reducing clutter*

4.1.1. Powers of Two Construction. Fig.4.1 illustrates the tile set of size $2n$, used in a powers of two construction. Attachment of S_1 to Seed is deterministic. The assembly halts only when the sequence $T_{n-1}, T_{n-2}, T_{n-3}, \dots T_2, T_1$ of attachments is achieved. The tiles R_i, $i = 1, 2 \dots, n-1$, each have g_n as their East pads and hence act as reset tiles. Each equiprobable choice is between a reset (addition of R_i) and progress towards completion (addition of T_i).

LEMMA 2. *Let X_i be the random variable equal to the number of tiles of type T_i in a terminal assembly. Then $E[X_{i-1}] = \frac{E[X_i]}{2}$ and $E[X_i] = 2^{i-1}$ for $i = 2, 3, \dots, n-1$. Let Y_i be the random variable equal to the number of tiles of type R_i in a final assembly. Then $E[Y_i] = E[X_i]$ for $i = 1, 2, 3, \dots, n-1$.*

Proof. Every time a tile of type T_i appears it is followed immediately after by either a tile of type T_{i-1} or R_{i-1} each with probability 1/2 for $i = 2, 3 \dots n-1$. So $E[Y_{i-1}] = E[X_{i-1}] = \frac{E[X_i]}{2}$. T_1 is the terminal tile and appears exactly once. Hence its expectation is $1 = 2^0$. Repeated application of the above geometric decrease property gives $E[X_i] = 2^{i-1}$ for $i = 1, 2, 3, \dots, n-1$. □

LEMMA 3. *Let L be the random variable equal to the length of the assembly for the powers of two construction (Fig. 4.1). Then $E[L] = 2^n$.*

Proof. $L = 2 + \sum_{i=1}^{n-1}(X_i + Y_i)$. Hence $E[L] = 2 + 2\sum_{i=1}^{n-1} E[X_i] = 2 + 2\sum_{i=1}^{n-1} 2^{i-1} = 2^n$. □

4.1.2. Extension to Arbitrary N. We extend the powers of two construction to all N by expressing N in binary, denoted by $B(N)$. For the i^{th} bits of $B(N)$ equal to 1, we have a power of two construction of expected length 2^i, using $2i$ tile types as in Section 4.1.1. 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. *Let L be the random variable equal to the length of the assembly described above. Then, $E[L] = N$. Thus, an assembly of expected length N can be constructed using $O(\log^2 N)$ tile types for any given $N \in \mathbb{N}$.*

Proof. As before, let $B(N)$ be the binary representation of N. Also, let $b(i)$ be a binary 0, 1 function which is equal to the i^{th} bit of $B(N)$. Now, $N = \sum_{i=0}^{\lfloor \log N \rfloor} b(i)2^i$. We have a power of two construction described in Section 4.1.1 of expected length 2^i, using $2i$ tile types, for each i for which $b(i) = 1$. From linearity of expectation, the expected length of the full assembly $E[L] = N$. The number of tile types used is upper bounded by $\sum_{i=0}^{\lfloor \log N \rfloor} b(i)(2i) \leq \sum_{i=0}^{\lfloor \log N \rfloor} 2i = O(\log^2 N)$. □

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 assemblies of length N in expectation for any given N using $\Theta(\log N)$ tile types. For powers of two, this construction reduces to the one in Section 4.1.1. Our construction for general N is more succinct than the one presented in Section 4.1.2. This new construction exploits the observation that the expected number of tiles of each type present in the powers of two construction decrease geometrically. We give an alternate binary encoding (see Li and Vitanyi (1997)) of 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 $\{1,2\}$ encoding of a non-zero natural number N is the N^{th} string in the lexicographic ordering of strings in $\{\mathbf{1},\mathbf{2}\}^+$. An equivalent characterization is 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 Lemma. Fig.4.2 is an example illustrating the construction for $N = 92$. For any given N, let N'' be the greatest even number less than N. For $N' = \frac{N''}{2}$, let $B(N') = b_{n-1}b_{n-2}\dots b_0$ be its $\{\mathbf{1},\mathbf{2}\}$-binary encoding of size n. For each bit b_i our construction has a progress tile complex T_i and a corresponding restart tile complex R_i of size b_i tiles each. Complexes T_i and R_i occurs X_i and Y_i times respectively. By an argument similar to Lemma 2, $E[X_i] = E[Y_i] = 2^{i-1}$. For even N, we deterministically prefix a single tile P to the West of the seed tile. For odd N we omit this prefix tile.

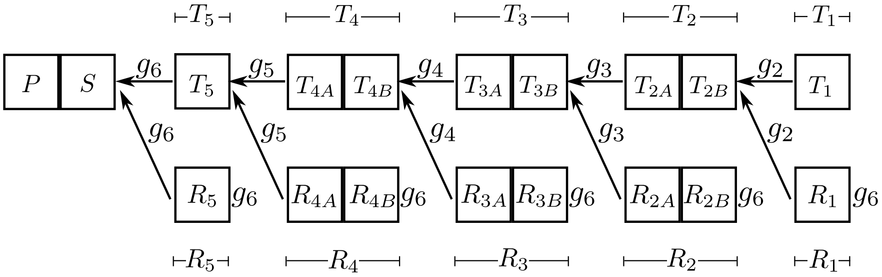

Fig. 4.2: *Tile binding diagram:* $N = 92; N'' = 90; N' = \frac{N''}{2} = 45 = (12221)_{alt2}$. *P is the prefix tile.*

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

Proof. Let X_i be the random variable equal to the number of times the complex T_i appears in the final assembly. Let Y_i be the random variable equal to the number of times the complex R_i appears in the final assembly. The length of the assembly L is given by

$$L = \sum_{i=0}^{n-2} (X_{i+1} + Y_{i+1}) b_i + (N + 1 \mod 2) + 1$$

and hence by linearity of expectation and Lemma 2,

$$E[L] = 2(\sum_{i=0}^{n-2} b_i 2^i) + (N + 1 \mod 2) + 1 = 2N' + (N + 1 \mod 2) + 1 = N$$

The number of tile types is $\Theta(n) = \Theta(\log N)$. □

5. Improving Tail Distributions of Linear Assemblies. In the previous section we achieved a linear assembly of expected length N using $O(\log N)$ tile types for all $N \in \mathbb{N}$. However, the tail bounds on the distribution of lengths achieved by tile sets of this construction is unclear. In this section we look at a general method to improve the tail distributions of linear assemblies at the expense of a linear factor increase in the tile complexity. We then give a concrete construction that uses $O(\log^3 N)$ tile types and prove that it has exponentially dropping tail distribution for infinitely many N.

5.1. Concatenating Independent Assemblies. 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. We can use the central limit theorem or Chernoff bounds (see Feller (1968); Motwani and Raghavan (1995)) for bounding the tail of this new distribution. Both the approaches are discussed below.

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, mutually disjoint versions of T, say $T_1, T_2, \ldots, T_k$. 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 be the random variable equal to the length of the assembly produced by this construction. This new multiset, T_{sh} of cardinality $f_{\text{sh}}(N) \leq kf(\lfloor \frac{N}{k} \rfloor) + k$ 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.

5.1.1. Central Limit Theorem Applied to Tail Distributions of Concatenated Assemblies. The central limit theorem gives:

$$\forall \delta \geq 0 : P(|L - N| \leq \delta\sigma) \to \Phi(\delta) \text{ as } k \to \infty,$$

where Φ is the probability density function of the standard normal distribution. Let ψ be the cumulative distribution function of the standard normal distribution. Thus,

$$P(|L - N| \geq \delta\sigma) \to 2(1 - \Phi(\delta)) \leq 2\psi(\delta)/\delta \leq \sqrt{2/\pi}(e^{-\delta^2/2}/\delta) \text{ as } k \to \infty.$$

Thus, we approach an exponentially decaying tail bound for large k, paying only a linear multiplicative increase in tile complexity.

5.1.2. Chernoff Bounds Applied to Tail Distributions of Concatenated Assemblies. Since T_{sh} is the concatenation of independent assemblies T_i, Chernoff bounds for sums of independent random variables gives

$$\forall \delta, t > 0 : P(L > (1+\delta)N) \leq (M(t)/e^{(1+\delta)\lfloor \frac{N}{k} \rfloor t})^k \text{ and}$$

$$\forall \delta > 0, t < 0 : P(L < (1-\delta)N) \leq (M(t)/e^{(1-\delta)\lfloor \frac{N}{k} \rfloor t})^k$$

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.

5.2. Linear assemblies of Expected Length N using $O(\log^3 N)$ tile types with Sharp Tail Bounds. We obtain a linear assembly of expected length N using $O(\log^3 N)$ tile types for all $N \in \mathbb{N}$. As before, we give a construction valid infinitely often, for $N = (n+1)2^n + 1$ for all $n \in \mathbb{N}$ and then later extend this to all $N \in \mathbb{N}$ by encoding N in terms of repeated units of our special construction. Next, we show how to use this construction to obtain assemblies which have sharp tail bounds on the distribution of lengths for infinitely many N.

5.2.1. Infinitely Often Construction using $O(\log^2 N)$ Tile Types. Fig.5.1 illustrates the tile set used to obtain a linear assembly of expected length $N = (n+1)2^n + 1$ for all $n \in \mathbb{N}$. The assembly halts only when the sequence of tiles $T_1, T_2, \ldots T_{n+1}$ attach. Either the West side of tile T_{i+1} or one of the unique restart sequences $B_{i,i+1}, B_{i,i+2}, \ldots B_{i,n+1}$ can attach to the East side of tile T_i for each i. Attachment of T_{i+1} is progress towards completion while the restart sequence sets the process back to step one. The restart sequences are unique in order to preserve the diagonal nature of the assembly. The process is akin to tossing a biased coin till a *head* appears. Each toss adds a linear chunk of $n+1$ tiles. A *tail* chunk is of the form $T_1, T_2 \ldots T_i B_{i,i+1}, B_{i,i+2}, \ldots B_{i,n+1}$. A *head* chunk is of the form $T_1, T_2 \ldots T_{n+1}$. *Head* chunk attaches with a probability of $\frac{1}{2^n}$, since each tile addition to the growing *head* chunk offers two equally likely possibilities of which exactly one is favorable. This is clearly a geometric process with parameter $\frac{1}{2^n}$ and so the expected number of chunks is 2^n. The size of each chunk $n+1$, giving a total assembly of length $N = (n+1)2^n + 1$ including the seed tile. The number of tile types used is $O(n^2) = O(\log^2 N)$.

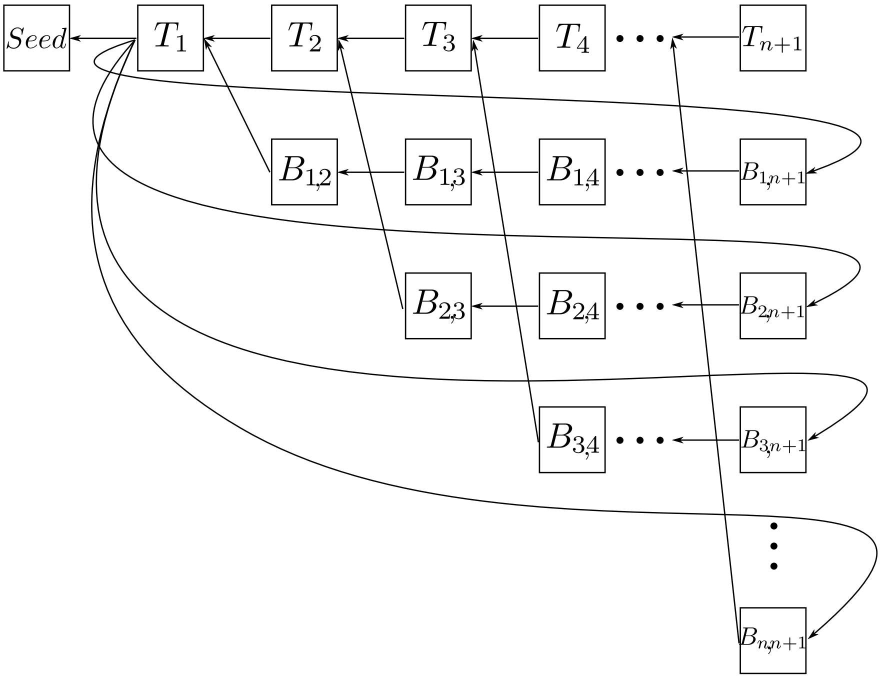

Fig. 5.1: Infinitely often construction using $O(\log^2 N)$ tile types

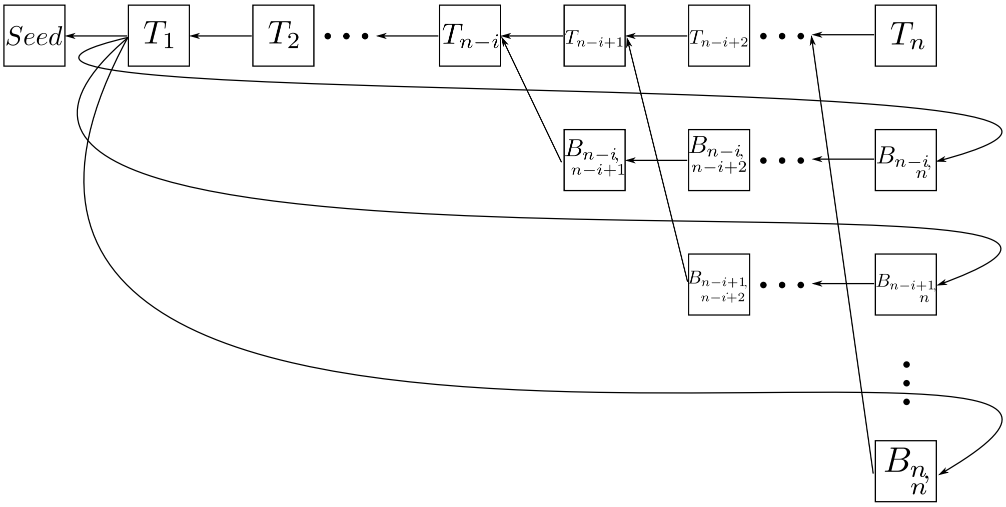

Fig. 5.2: Chains of expected length $n2^i$ for all $i \in \{0, 1, \ldots, n-1\}$.

5.2.2. Extension to Arbitrary N**.** We extend the above infinitely often construction to all N by the following encoding scheme. First we will show how to construct chains of expected length $n2^i$ for all $i \in \{0, 1, \ldots, n-1\}$. Then, we express N as chunks of $n2^i$ which we put together to obtain an assembly close to N. The small remainder is constructed using unique tiles.

Fig.5.2 illustrates chains of expected length $n2^i$, excluding the seed tile, for all $i \in \mathbb{N}$. The construction is a small modification of our earlier infinitely often construction. The *head* and *tail* chunks are of size n but the stochastic process involves only the tiles T_{n-i} to T_n and restart sequences $n-i+1$ and beyond. This leads to a smaller bias of 2^i and the expected length of such assemblies is $n2^i$, excluding the seed tile. This construction uses $O(n^2)$ tile types.

Now let $N = a\lceil \log N \rceil + b$ where a, b are non-negative integers and $b < \lceil \log N \rceil$. By expressing a in its binary representation, we can achieve constructions of size $\lceil \log N \rceil 2^i$ that when put together deterministically, give $a\lceil \log N \rceil$ in expectation by the property of linearity. Let $B(a)$ be the binary representation of a. Let $b(i)$ be a binary function equal to the i^{th} bit of $B(a)$. Then, $a = \sum_{i=0}^{\lfloor \log a \rfloor} b(i)2^i$ and so $a\lceil \log N \rceil = \sum_{i=0}^{\lfloor \log a \rfloor} b(i)(\lceil \log N \rceil 2^i)$. We create subassemblies of expected length $\lceil \log N \rceil 2^i$ for each i such that $b(i) = 1$ using the construction described above. Putting these assemblies together gives a linear assembly of expected length $a\lceil \log N \rceil$. For the remainder $b < \log N$ we use unique tiles. Thus the expected length achieved by this construction is N by linearity of expectation. This length excludes the single seed tile, which can be adjusted by programming for linear assemblies of length $N - 1$.

Each of the subassemblies requires $\lceil \log N \rceil - i + \frac{i^2}{2}$ tile types. Thus, the total number of tile types to form $a\lceil \log N \rceil$ is at most $\sum_{i=0}^{\lfloor \log a \rfloor} \left(\lceil \log N \rceil - i + \frac{i^2}{2} \right) = O(\log^3 N)$. The remainder b uses at most $\lceil \log N \rceil$ unique tiles. Thus, the total number of tile types is $O(\log^3 N)$.

5.2.3. Tail Bound for Linear Assemblies of Expected Length N using $O(\log^3 N)$ Tile Types. Here we show how to obtain linear assemblies with sharp tail bounds on distribution of lengths. Recall the earlier discussion where the assembly was thought of as a sequence of independent coin tosses, each adding an n length chunk to the assembly. Assembly process halts when the first head chunk is realized. Unfortunately, this process results in a geometric random process which does not have a sharp left tail. To obtain good tail bounds, we repeat the assembly process r times to realize a negative binomial random process. We do this by deterministically concatenating r distinct versions of the assembly together. This is done by allowing the West side of the starting tile of the $i + 1^{\text{th}}$ version to bind to the East side of the final tile of the i^{th} version. As earlier, the analysis below excludes the seed tile, which can be adjusted by programming tile sets to construct linear assemblies of expected length $N - 1$.

Suppose we are given a target length $N = rn2^i$ where r, n, i are positive integers less than $\lceil \log N \rceil$. We build r versions of the assembly described earlier, each with expected length $n2^i$ using at most $O(rn^2) = O(\log^3 N)$ tile types. This gives a linear assembly of expected length N. We prove in the theorem below that such assemblies have exponentially dropping tail bounds.

Theorem 3. *Let L be the random variable equal to the length of the above linear assembly, with $E[L] = N$. Then for all $0 < \alpha$ and $0 < \beta < 1$*

$$Pr[L > (1 + \alpha)N] < \left(\frac{1 + \alpha}{e^{\alpha}} \right)^r \text{ and } Pr[L \leq (1 - \beta)N] \leq \left(\frac{1 - \beta}{e^{-\beta}} \right)^r$$

for infinitely many N.

Proof. Let X be the random variable whose value is the number of n-length chunks in the final linear chain. Then X is negative binomial random variable with parameters r (number of successes) and $p = \frac{1}{2^i}$ (the probability of success). Note that $L = nX$ and $E[L] = nE[X]$.

Let $\mathcal{N}(r, p)$ be a negative binomial random variable with parameters r (number of successes) and p (probability of a success). Let $\mathcal{B}(M, p)$ be a binomial random variable with parameters M (number of Bernoulli trials) and p (probability of a success). It is well known that

$$\Pr[\mathcal{N}(r, p) > M] = \Pr[\mathcal{B}(M, p) < r] \text{ and } \Pr[\mathcal{N}(r, p) \leq M] = \Pr[\mathcal{B}(M, p) \geq r]$$

Thus, we can use Chernoff bounds derived for the binomial distribution to obtain tail bounds for the negative binomial distribution using the aformentioned relationship. Mitzenmacher and Upfal (2005) derive the following Chernoff bounds for the binomial random variable Y with mean $E[Y] = \mu$ and for $\delta > 0$:

$$\Pr[Y \geq (1 + \delta)\mu] \leq \left(\frac{e^{\delta}}{(1 + \delta)^{(1+\delta)}} \right)^{\mu} \text{ and } \Pr[Y < (1 - \delta)\mu] < \left(\frac{e^{-\delta}}{(1 - \delta)^{(1-\delta)}} \right)^{\mu}$$

Recall that X is a negative binomial random variable with parameters r (number of successes) and $p = \frac{1}{2^i}$ (the probability of success). Then $E[X] = \frac{r}{p}$. By the result stated previously, $\Pr[X > (1+\alpha)\frac{r}{p}] = \Pr[Y < r]$ where Y is a binomial random variable with parameters $(1+\alpha)\frac{r}{p}$ (number of Bernoulli trials) and p (probability of success). Thus $\mu = E[Y] = (1+\alpha)\frac{r}{p}p = (1+\alpha)r$ and so $r = \frac{\mu}{1+\alpha} = (1-\delta)\mu$ where $\delta = \frac{\alpha}{1+\alpha}$. Thus $\Pr[Y < r] = \Pr[Y < (1-\delta)\mu] < \left(\frac{e^{-\delta}}{(1-\delta)^{(1-\delta)}}\right)^{\mu} = \left(\frac{e^{-\delta}}{(1-\delta)^{(1-\delta)}}\right)^{(1+\alpha)r} = \left(\frac{1+\alpha}{e^{\alpha}}\right)^{r}$. Hence $\Pr[X > (1+\alpha)E[X]] < \left(\frac{1+\alpha}{e^{\alpha}}\right)^{r}$ and thus $\Pr[L > (1+\alpha)N] < \left(\frac{1+\alpha}{e^{\alpha}}\right)^{r}$. Note that by Taylor series, $\frac{1+\alpha}{e^{\alpha}} < 1$ for any positive α.

By a similar argument, one gets $\Pr[L \leq (1-\beta)N] \leq \left(\frac{1-\beta}{e^{-\beta}}\right)^{r}$. □

These tail bounds further drop exponentially with r. Recall that to obtain a tile complexity of $O(\log^3 N)$, we need to choose some $r < \lceil \log N \rceil$. For large N, r becomes large enough to obtain sharp tail bounds on the length of the linear assemblies.

Thus, we have demonstrated how to obtain linear assemblies with sharp tail bounds on distribution of lengths for infinitely many N using $O(\log^3 N)$ tile types.

6. Lower Bounds on the Tile Complexity of Linear Assemblies of Expected Length N in PTAM. 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.

Any standard PTAM linear tiling multiset with cardinality n that produces linear assemblies of greatest (finite) expected length is termed *n-optimal* or simply *optimal.*

Lemma 5. *Optimal linear tiling multisets must contain exactly one capping tile.*

Proof. Suppose an optimal multiset has multiple capping tiles $term_1, \ldots, term_k$. 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. □

The following technical lemma will be needed in Theorem 4.

Lemma 6. *Let $z, k_i \in \mathbb{Z}_{>0}$ for $i = 1, 2, \ldots, z$. If $\sum_{i=1}^{z} k_i = m$ then the maximum value of $\prod_{i=1}^{z}(k_i + 1)$ is* 2^m

Proof. Applying the inequality of arithmetic and geometric means,

$$\prod_{i=1}^{z}(k_i+1) \leq \left(\frac{\sum_{i=1}^{z}(k_i+1)}{z}\right)^{z} = \left(\frac{m+z}{z}\right)^{z} = \left(1+\frac{m}{z}\right)^{z} = \left(\left(1+\frac{m}{z}\right)^{\frac{z}{m}}\right)^{m}$$

Note that $\frac{m}{z} \geq 1$. The function $(1+x)^{\frac{1}{x}}$ is strictly decreasing and hence the maximum value of 2^m for the above expression is obtained when $m = z$. The reader may verify that the maximum value is indeed attained by substituting $k_i = 1$ for all i. □

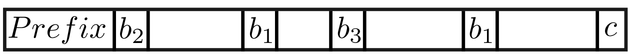

Fig. 6.1: $\mathbb{T}$ *split into prefix and intermediates.*

Theorem 4. *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 $O(2^n)$ unless the expected length is infinite. This implies our result via the contrapositive. Recall that the multiplicity of tiles in the multiset count distinctly towards tile complexity. 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 with capping tile c. Let L be the random variable equal to the length of the linear assembly produced by $\mathbb{T}$ and so $E[L] = \Psi_n$. For every terminal assembly $st_1t_2\dots c$ of $\mathbb{T}$ we define a *run* as the sequence of pad types

$$\mathrm{pad}_{\mathrm{East}}(s), \mathrm{pad}_{\mathrm{East}}(t_1), \mathrm{pad}_{\mathrm{East}}(t_2), \dots, \mathrm{pad}_{\mathrm{East}}(c)$$

Let $\lambda = \mathrm{pad}_{\mathrm{West}}(c)$ be the pad type appearing on the West side of the capping tile. We define $\Lambda = \{c, b_1, b_2, \dots, b_{k_1-1}\} \subset T$ as the multiset of k_1 tiles with λ as their West pad ($0 < k_1 < n$). Pad type λ might occur at many positions in a run. The subsequence of a run from $\mathrm{pad}_{\mathrm{East}}(s)$ to the first occurrence of λ is termed as the *prefix* of the run. The subsequence of a run that starts with $\mathrm{pad}_{\mathrm{East}}(b_i)$ and ends in λ with no occurrence of λ within is termed the i^{th} intermediate of the run (Figure 6.1).

Define the following random variables corresponding to a run: L_P equal to the length of the prefix, L_i equal to the length of the i^{th} intermediate and r_i equal to number of times the i^{th} intermediate occurs in the run. By definition, the length of the assembly $L = L_P + \sum_{i=1}^{k_1-1}(r_iL_i) + 1$. Note that r_i and L_i are independent random variables because of the memoryless property of linear assemblies. That is, the length of an intermediate is independent of the number of times that intermediate occurs in a run. Thus, by linearity of expectation and independence we get,

$$\Psi_n = E[L] = E[L_P] + \sum_{i=1}^{k_1-1} E[r_iL_i] + 1 = E[L_P] + \sum_{i=1}^{k_1-1} E[r_i]E[L_i] + 1$$

Note that the number of times the i^{th} intermediate occurs in a run equals the number of times tile b_i attaches to the assembly. The tiles in Λ can attach with equal probability $\frac{1}{k_1}$ to any tile with λ as its East pad. If the capping tile attaches, the run stops, else it continues. The process is akin to rolling a k_1 sided die till k_1 appears and counting the expected number of times a certain roll is achieved and hence $E[r_i] = 1$ for $i = 1, 2, \dots, k_1 - 1$.

We will show that $E[L_P]$ and $E[L_i]$ are at most Ψ_{n-k_1} by simulating the subassemblies 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 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 $E[L_P] \le \Psi_{n-k_1}$ by definition. The i^{th} intermediate is simulated by a tile multiset T_i of cardinality $n - k_1$ obtained from T by (i) dropping the tiles in Λ from T and (ii) replacing the seed tile s by the tile $(\phi, \mathrm{pad}_{\mathrm{East}}(b_i))$. Again, we observe that there is a run of $\mathbb{T}_i$ for every possible i^{th} intermediate and vice-versa, with the same probabilities of formation. Thus $E[L_i] \le \Psi_{n-k_1}$. Substituting, we get the inequality,

$$\Psi_n = E[L_P] + \sum_{i=1}^{k_1-1} E[r_i]E[L_i] + 1 \le k_1\Psi_{n-k_1} + 1 \le (k_1 + 1)\Psi_{n-k_1}$$

In the next level of recursion, we drop $k_2 > 0$ tiles to get $\Psi_n \le (k_1+1)\Psi_{n-k_1} \le (k_1+1)(k_2+1)\Psi_{n-k_1-k_2}$. In general, we drop k_i tiles in the i^{th} level of recursion to get $\Psi_n \le \prod_{j=1}^{i}(k_j + 1)\Psi_{n-\sum_{j=1}^{i} k_j}$. 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 $\Psi_n \le \prod_{i=1}^{z}(k_i + 1)$ 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} (Lemma 6). Hence $\Psi_n \le O(2^n)$. □

7. 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.7.1). To ensure that tiles align fully and are not offset, each pad on a side of a tile is 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.

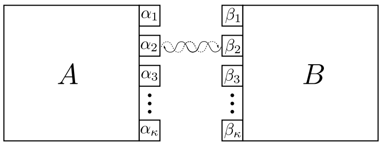

Fig. 7.1: *κ-pad tiles A & B.*

7.1. 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[3] (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 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.7.1).

With these modifications, *diagonal*, *uni-seeded* and *haltable* linear tiling systems and self-assembly of κ-pad tiles are defined as in Section 2 and Section 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 7.3, adjacent tiles that bind have exactly one match among corresponding pads.

7.2. 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, developed by Winfree et al. (1998), which is one of the early demonstrations of DNA motifs that self-assemble into two dimensional lattices, can serve as a 2-pad tile with slight modifications to ensure that tiles align correctly as they attach. Other similar motifs that also self-assemble into two dimensional lattices, like the TX developed by Mao et al. (2000) and the DDX developed by Reishus et al. (2005), can serve as multipad systems with similar modifications.

These motifs can be easily modified to self-assemble in one dimension, as a linear structure. On a much larger scale, origami techniques developed by Rothemund (2006) 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. However, this drawback may be somewhat mitigated by letting multiple pads act as a single virtual pad.

7.3. Linear Assemblies of Expected Length N using $\Theta_{i.o}\left(\frac{\log N}{\log\log N}\right)$ 2-pad Tile Types.

In this section we present a 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 tiles, i.e., arbitrary long fixed length assemblies of expected length N using $\Theta\left(\frac{\log N}{\log\log N}\right)$ 2-pad tiles. Fig.7.2 illustrates the tile set used in our construction. $Q_1, Q_2, Q_3 \ldots Q_{n-1}$ are tiles with multiplicity 1. R is a tile type with multiplicity $n-1$, drawn as $R_1, \ldots, R_{n-1}$ in Fig.7.2. Q_{i-1} can attach to Q_i's East side via the upper pad g_{i-1}. For $i \in \{1, 2, \ldots, n-1\}$, $R_1, R_2, \ldots, R_{n-1}$ can attach to Q_i's East side via the lower pad b. Q_n is the capping

[3] 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.

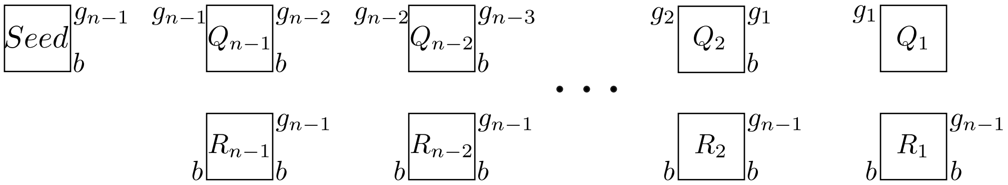

Fig. 7.2: *Pad binding diagram for linear tiling system using* $\Theta_{i.o}(\frac{\log N}{\log\log N})$ *2-pad tile types. Arrows are omitted to reduce clutter. Absent pads are* ϕ.

tile and *Seed* is the seed tile. The assembly halts iff the consecutive sequence $Q_{n-1}, Q_{n-2}, \ldots, Q_1$ occurs. At each stage, the assembly can restart by the attachment of one of the $n-1$ bridge tiles R_i. The number of tile types is $2(n-1)+1 = \Theta(n)$.

Lemma 7. *Let* X_i *be the random variable equal to the number of tiles of type* Q_i *in a final assembly. Then* $E[X_i] = nE[X_{i-1}]$ *for* $i = 2, 3, \ldots, n-1$ *and* $E[X_i] = n^{i-1}$ *for* $i = 1, 2, 3, \ldots, n-1$. *Let* Y *be the random variable equal to the number of tiles of type* R *in a final assembly. Then* $E[Y] = (n-1)\sum_{i=1}^{n-1} E[X_i]$.

Proof. Every time a tile of type Q_i appears the probability of a tile of type Q_{i-1} attaching is $1/n$ for $i = 2, 3 \ldots n-1$. So $E[X_i] = nE[X_{i-1}]$. Q_1 is the terminal tile and appears exactly once. Hence its expectation is $1 = 2^0$. Repeated application of the above geometric decrease property gives $E[X_i] = n^{i-1}$ for $i = 1, 2, 3, \ldots, n-1$. Also, every time a tile of type Q_i appears, one of the bridge tiles R is $n-1$ times as likely to appear as a tile of type Q_{i-1}. Hence $E[Y] = (n-1)\sum_{i=1}^{n-1} E[X_i]$ □

Theorem 5. *Let* L *be the random variable that equals the length of the tile system illustrated in Fig.7.2. Then* $E[L] = N = \Theta(n^{n-1})$ *using* $\Theta\left(\frac{\log N}{\log\log N}\right) = \Theta(n)$ *2-pad tile types.*

Proof. $L = 1 + Y + \sum_{i=1}^{n-1} X_i$. Taking expectations and applying Lemma 7,

$$E[X] = 1 + E[Y] + \sum_{i=1}^{n-1} E[X_i] = 1 + (n-1)\sum_{i=1}^{n-1} n^{i-1} + \sum_{i=1}^{n-1} n^{i-1} = \frac{n^n - 1}{n-1} = \Theta(n^{n-1})$$

The number of tile types used is $\Theta(n) = \Theta\left(\frac{\log N}{\log\log N}\right)$. □

7.4. 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 $\Omega\left(\frac{\log N}{\log\log N}\right)$.

Theorem 6. *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* $\Omega\left(\frac{\log N}{\log\log N}\right)$.

Proof. As in the Theorem 4, we will show that any κ-pad standard PTAM system with tile multiset of cardinality n has expected length of assembly at most $O\left(n^{n+1}\right)$ (or else infinite) and this implies our result via the contrapositive.

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 6. Let L be the random variable equal to the length of linear assembly produced by $\mathbb{T}$. For the sake of clarity in the rest of this proof we distinguish each tile in T by a *face label* that does not play any role in binding probabilities. Thus if a tile has multiplicity greater than 1 we distinguish the multiple copies via their distinct *face labels.*

Consider a non-terminal produced configuration. Suppose the last (East-most) tile attached was v_0 (Note: v_0 is not the capping tile). Since the assembly is haltable, there exists a finite sequence of tile additions that halt the assembly and no two tiles in the sequence are identical. Suppose $v_0, v_1, .., v_k$ is some such sequence where v_k is the capping tile. Note that $1 \leq k < n$ since each tile in this sequence is distinct. There are at most n tiles competing for attachment at any stage of assembly and every possible attachment (i.e. non-zero probability attachment) is equally likely. Hence each tile attachment in this sequence has a probability of attachment at least $\frac{1}{n}$. Thus, the probability of the assembly halting after k attachments is at least $1/n^k > 1/n^n$ and the number of tiles added is $k < n$. Thus, the assembly process can be thought of as a sequence of Bernoulli trials until success is obtained. Each *failed* trial corresponds to a sequence of n attachments not containing the capping tile. A *successful* trial corresponds to a sequence of $k < n$ attachments ending with the capping tile. The probability of success is at least $\frac{1}{n^n}$ and hence the expected number of trials till success is at most n^n. Each trial adds at most n tiles and so the expectation of the assembly is upper bounded by $n \times n^n = n^{n+1}$.

Thus the expected length of an assembly of any κ-pad standard linear PTAM system with tile multiset of cardinality n is at most $O\left(n^{n+1}\right)$ which implies a lower bound of $\Omega\left(\frac{\log N}{\log \log N}\right)$. □

8. Conclusions and Future Work. Fixed length linear structures are important components for engineering DNA nanostructures. This work proposes ways to construct linear assemblies in a tiling model using very few tile types by using stochastic, non-deterministic behavior. We extended TAM to PTAM to take advantage of inherent probabilistic behavior in many self-assembled systems. A restriction to standard PTAM was considered in this article. Prior work in DNA self-assembly strongly suggests that standard PTAM can be realized in the laboratory. We showed various non-trivial probabilistic constructions for forming linear assemblies in PTAM with tile sets of sub-linear cardinality, using techniques that differ considerably from existing assembly techniques. In particular, for any given N, we demonstrated linear assemblies of expected length N with a tile set of cardinality $\Theta(\log N)$ using one pad per side of each tile. We proved a lower bound of $\Omega(\log N)$ for each N on the tile complexity of linear assemblies of expected length N in standard PTAM systems using one pad per side of each tile. We further demonstrated how linear assemblies can be modified to produce assemblies with sharp tail bounds on distribution of lengths by concatenating various assemblies together. In particular, we showed that for infinitely many N we can get linear assemblies with exponentially dropping tail distributions using $O(\log^3 N)$ tile types.

We also proposed a simple extension to PTAM called κ-pad systems in which we associate κ pads with each side of a tile. This gives linear assemblies of expected length N with a 2-pad tile set of cardinality $\Theta(\frac{\log N}{\log \log N})$ for infinitely many N. We showed that we cannot get smaller tile complexity by proving a lower bound of $\Omega(\frac{\log N}{\log \log N})$ for all N on the cardinality of the κ-pad tile multiset required to form linear assemblies of expected length N in standard κ-pad PTAM systems. The techniques that we used for deriving these tile complexity lower bounds are notable as they differ from traditional Kolmogorov complexity based information theoretic methods used for lower bounds on tile complexity. Also, Kolmogorov complexity based lower bounds do not preclude the possibility of achieving assemblies of very small tile multiset cardinality for infinitely many N. In contrast, our lower bounds are stronger as they hold for every N. We also answered the question of what is the longest finite linear assemblies one can construct using given cardinality tile multisets in PTAM and κ-pad PTAM. Thus, for linear assembly systems, we have shown that stochastic behavior at the level of tiles can be exploited to get large improvements in tile complexity at a small expense of precision in length.

Self-assembled DNA systems are error prone (see Winfree and Bekbolatov (2003); Chen and Goel (2004); Reif et al. (2004)). Two particular kind of errors that affect assemblies in TAM are spurious nucleation and pad mismatch. The Probabilistic Tile Assembly Model is also affected by these as it is an extension of TAM and makes some of the same key assumptions, but pad mismatch errors can be modeled with minimal changes to the PTAM model due to its stochastic nature. An immediate question is how to implement schemes for error correction, reduction and avoidance in PTAM. In particular, how do we construct robust linear assemblies in the presence of the aforementioned errors. Can error correction, reduction and avoidance schemes leverage the stochastic nature of PTAM to produce robust assemblies?

Adleman et al. (2001a) studied the notion of running time of an assembly for TAM, a notion that is extendable to PTAM. Since PTAM systems encode concentrations in their tile multiset, their running times are implicitly specified. Note that it takes time $\Omega(N)$ to assemble an N length linear assembly in PTAM. The linear tile multiset of Becker et al. (2006) has an optimal running time of $\Theta(N)$ but has suboptimal tile multiset complexity $\Omega(N)$. In comparison, the linear tile system we presented in Section 4.2 has optimal tile multiset cardinality of $\Theta(\log N)$ but suboptimal running time of $\Omega(N \log N)$ (proof follows directly from the observation that the system is equimolar). For standard PTAM systems, can a linear assembly obtain both optimal running time $O(N)$ and optimal tile complexity $O(\log N)$ (defined as the cardinality of the tile multiset)?

A more general model of tiling, as proposed by Aggarwal et al. (2004), allows preformed assemblies consisting of multiple tiles (called *supertiles*) to attach to each other and form a larger supertile. The assembly time in such models has been considered in Chen and Doty (2012) and Adleman et al. (2001b). What is the appropriate notion of assembly time for PTAM systems that allow attachment of supertiles and how would the assembly time of the systems described in Section 4 compare to the optimal assembly times under the supertile attachment assumption? Also, the tail bounds derived in Section 5.2.3 only apply for infinitely many N and require $O(\log^3 N)$ tile types. Can we obtain tail bounds for all sufficiently large N? Can we reduce the tile complexity required to below $O(\log^3 N)$? Finally, it would be interesting to perform experimental verification of our proposed systems. As experimental demonstration would involve tile concentrations, PTAM must be expanded to accommodate finite precision concentration programming. In that expanded setting, tradeoffs between tile complexity and number of bits allowed to specify tile concentrations can be studied. Doty (2010) has studied a closely related question for assembling squares in the concentration programming model.

Acknowledgements. We thank anonymous reviewers for comments and suggestions that improve the organization and general readability of the paper. In particular, we thank the reviewers for a suggestion that simplifies our constructions in Figures 4.1 and 4.2 and the corresponding analysis. An earlier version of Theorem 6 was replaced with the current simpler argument at the suggestion of a reviewer. This work was supported by NSF grants CCF-1141847 and CCF-0829797. An earlier version of this paper appeared as Chandran et al. (2009).

References.

Adleman, L. (2000). Towards a Mathematical Theory of Self-assembly. Technical report, University of Southern California.

Adleman, L., Cheng, Q., Goel, A., and Huang, M.-D. (2001a). Running Time and Program Size for Self-Assembled Squares. *Symposium on Theory of Computing*, pages 740–748.

Adleman, L., Cheng, Q., Goel, A., Huang, M.-D., and Wasserman, H. (2001b). Linear Self-assemblies: Equilibria, Entropy, and Convergence Rates. *ICDEA*.

Aggarwal, G., Cheng, Q., Goldwasser, M., Kao, M.-Y., de Espanes, P. M., and Schweller, R. (2005). Complexities for Generalized Models of Self-Assembly. *SIAM Journal on Computing*, 34(6):1493–1515.

Aggarwal, G., Goldwasser, M., Kao, M.-Y., and Schweller, R. (2004). Complexities for Generalized Models of Self-Assembly. *Symposium on Discrete Algorithms*, pages 880–889.

Andersen, E., Dong, M., Nielsen, M., Jahn, K., Subramani, R., Mamdouh, W., Golas, M., Sander, B., Stark, H., Oliveira, C., Pedersen, J. S., Birkedal, V., Besenbacher, F., Gothelf, K., and Kjems, J. (2009). Self-assembly of a Nanoscale DNA Box with a Controllable Lid. *Nature*, 459(7243):73–76.

Barish, R., Rothemund, P., and Winfree, E. (2005). Two Computational Primitives for Algorithmic Self-Assembly: Copying and Counting. *Nano Letters*, 5:2586–2592.

Becker, F., Rapaport, I., and Remila, E. (2006). Self-assemblying Classes of Shapes with a Minimum Number of Tiles, and in Optimal Time. *Foundations of Software Technology and Theoretical Computer Science*, pages 45–56.

Berger, R. (1966). The Undecidability of the Domino Problem. *Memoirs of American Mathematical Society*, 66:1–72.

Bryans, N., Chiniforooshan, E., Doty, D., Kari, L., and Seki, S. (2011). The Power of Nondeterminism in Self-Assembly. *SODA*, pages 590–602.

Chandran, H., Gopalkrishnan, N., and Reif, J. (2009). The Tile Complexity of Linear Assemblies. *International Colloquium on Automata, Languages and Programming*, pages 235–253.

Chen, H.-L. and Doty, D. (2012). Parallelism and Time in Hierarchical Self-assembly. *SODA*, pages 1163–1182.

Chen, H.-L. and Goel, A. (2004). Error Free Self-assembly Using Error Prone Tiles. *DNA Computing*, pages 62–75.

Chen, H.-L., Schulman, R., Goel, A., and Winfree, E. (2007). Reducing Facet Nucleation during Algorithmic Self-Assembly. *Nano Letters*, 7:2913–2919.

Demaine, E., Demaine, M., Fekete, S., Ishaque, M., Rafalin, E., Schweller, R., and Souvaine, D. (2007). Staged Self-assembly: Nanomanufacture of Arbitrary Shapes with $O(1)$ Glues. *DNA Computing*, pages 1–14.

Demaine, E., Demaine, M., Fekete, S., Ishaque, M., Rafalin, E., Schweller, R., and Souvaine, D. (2008). Staged Self-assembly: Nanomanufacture of Arbitrary Shapes with O(1) Glues. *Natural Computing*, 7(3):347–370.

Dietz, H., Douglas, S., and Shih, W. (2009). Folding DNA into Twisted and Curved Nanoscale Shapes. *Science*, 325(5941):725–730.

Dirks, R. and Pierce, N. (2004). Triggered Amplification by Hybridization Chain Reaction. *Proceedings of the National Academy of Sciences of the United States of America*, 101(43):15275–15278.

Doty, D. (2009). Randomized Self-Assembly for Exact Shapes. *Foundations of Computer Science*.

Doty, D. (2010). Randomized Self-Assembly for Exact Shapes. *SIAM Journal on Computing*, 39(8):3521–3552.

Doty, D., Patitz, M., and Summers, S. (2011). Limitations of Self-assembly at Temperature 1. *Theorectical Computer Science*, 412(1-2):145–158.

Douglas, S., Dietz, H., Liedl, T., Hogberg, B., Graf, F., and Shih, W. (2009). Self-assembly of DNA into Nanoscale Three-dimensional Shapes. *Nature*, 459(7245):414–418.

Feller, W. (1968). *An Introduction to Probability Theory and Its Applications.* Wiley.

Garey, M. and Johnson, D. (1981). *Computers and Intractability: A Guide to the Theory of NP-Completeness.* W. H. Freeman.

Gillespie, D. (1977). Exact Stochastic Simulation of Coupled Chemical Reactions. *The Journal of Physical Chemistry*, 81:2340–2361.

Kao, M.-Y. and Schweller, R. (2006). Reducing Tile Complexity for Self-assembly through Temperature Programming. *Symposium on Discrete Algorithms*, pages 571–580.

Kao, M.-Y. and Schweller, R. (2008). Randomized Self-Assembly for Approximate Shapes. *International Colloquium on Automata, Languages and Programming*, pages 370–384.

Lagoudakis, M. and LaBean, T. (1998). 2D DNA Self-Assembly for Satisfiability. *DIMACS Workshop on DNA Based Computers.*

Lewis, H. and Papadimitriou, C. (1981). *Elements of the Theory of Computation.* Prentice Hall.

Li, M. and Vitanyi, P. (1997). *An Introduction to Kolmogorov Complexity and Its Applications.* Springer.

Mao, C., Labean, T., Reif, J., and Seeman, N. (2000). Logical Computation Using Algorithmic Self-assembly of DNA Triple-crossover Molecules. *Nature*, 407:493–496.

Mitzenmacher, M. and Upfal, E. (2005). *Probability and Computing: Randomized Algorithms and Probabilistic Analysis.* Cambridge University Press.

Motwani, R. and Raghavan, P. (1995). *Randomized Algorithms.* Cambridge University Press.

Park, S.-H., Yin, P., Liu, Y., Reif, J., LaBean, T., and Yan, H. (2005). Programmable DNA Self-assemblies for Nanoscale Organization of Ligands and Proteins. *Nano Letters*, 5:729–733.

Reif, J., Sahu, S., and Yin, P. (2004). Compact Error-Resilient Computational DNA Tiling Assemblies. *DNA Computing*, pages 293–307.

Reishus, D., Shaw, B., Brun, Y., Chelyapov, N., and Adleman, L. (2005). Self-Assembly of DNA Double-Double Crossover Complexes into High-Density, Doubly Connected, Planar Structures. *Journal of the American Chemical Society*, 127(50):17590–17591.

Robinson, R. (1971). Undecidability and Nonperiodicity for Tilings of the Plane. *Inventiones Mathematicae*, 12:177–209.

Rothemund, P. (2006). Folding DNA to Create Nanoscale Shapes and Patterns. *Nature*, 440:297–302.

Rothemund, P., Papadakis, N., and Winfree, E. (2004). Algorithmic Self-Assembly of DNA Sierpinski Triangles. *PLoS Biology*, 2.

Rothemund, P. and Winfree, E. (2000). The Program-Size Complexity of Self-Assembled Squares. *Symposium on Theory of Computing*, pages 459–468.

Schulman, R., Lee, S., Papadakis, N., and Winfree, E. (2004). One Dimensional Boundaries for DNA Tile Self-Assembly. *DNA Computing*, pages 108–125.

Schulman, R. and Winfree, E. (2007). Synthesis of Crystals with a Programmable Kinetic Barrier to Nucleation. *Proceedings of the National Academy of Sciences of the United States of America*, 104:15236–15241.

Soloveichik, D. and Winfree, E. (2007). Complexity of Self-Assembled Shapes. *SIAM Journal of Computing*, 36:1544–1569.

Wang, H. (1961). Proving Theorems by Pattern Recognition II. *Bell Systems Technical Journal.*

Winfree, E. (1995). On the Computational Power of DNA Annealing and Ligation. *DNA Based Computers,DIMACS.*

Winfree, E. (1998a). *Algorithmic Self-Assembly of DNA*. PhD thesis, California Institute of Technology.

Winfree, E. (1998b). Simulations of Computing by Self-Assembly. Technical report, California Institute of Technology.

Winfree, E. and Bekbolatov, R. (2003). Proofreading Tile Sets: Error Correction for Algorithmic Self-Assembly. *DNA Computing*, pages 126–144.

Winfree, E., Liu, F., Wenzler, L., and Seeman, N. (1998). Design and Self-assembly of Two-dimensional DNA Crystals. *Nature*, 394:539–544.

Yan, H., Feng, L., LaBean, T., and Reif, J. (2003). Parallel Molecular Computation of Pair-Wise XOR using DNA String Tile. *Journal of the American Chemical Society*, 125.

Yin, P., Choi, H., Calvert, C., and Pierce, N. (2008). Programming Biomolecular Self-assembly Pathways. *Nature*, 451(7176):318–322.

Zhang, D., Turberfield, A., Yurke, B., and Winfree, E. (2007). Engineering Entropy-Driven Reactions and Networks Catalyzed by DNA. *Science*, 318:1121–1125.

Zheng, J., Birktoft, J., Chen, Y., Wang, T., Sha, R., Constantinou, P., Ginell, S., Mao, C., and Nadrian (2009). From Molecular to Macroscopic via the Rational Design of a Self-assembled 3D DNA Crystal. *Nature*, 461(7260):74–78.