Non-Determinism Reduces Construction Time in Active Self-Assembly Using an Insertion Primitive*

Benjamin Hescott[1], Caleb Malchik[2], and Andrew Winslow[3]

[1] Northeastern University, Boston, MA, USA
b.hescott@northeastern.edu
[2] Yale University, New Haven, CT, USA
caleb.malchik@yale.edu
[3] University of Texas Rio Grande Valley, Edinburg, TX, USA
andrew.winslow@utrgv.edu

Abstract. We consider efficient construction of DNA-based polymers in a model introduced by Dabby and Chen (SODA 2013) called *insertion systems*, where monomers insert themselves into the middle of a growing linear polymer. Specifically, we describe a new family of non-deterministic insertion systems that construct length-n polymers in $\Theta(\log^{3/2}(n))$ expected time, breaking the lower bound of $\Omega(\log^{5/3}(n))$ for deterministic construction. We also prove that this time is optimal for systems constructing finite polymers, and that the $\Theta(\log(n))$ monomer types used in the construction is optimal for this time.

1 Introduction

We study a theoretical model of DNA-based *algorithmic self-assembly* introduced by Dabby and Chen [6], in which simple particles called *monomers* aggregate to form long, complex chains called *polymers* via individual monomers asynchronously inserting themselves between adjacent monomers. This model shares similarities with other *active*[4] self-assembly models, e.g. the graph grammars of Klavins et al. [14, 15] and the *nubots* model of Woods et al. [2, 3, 19, 4], where structures undergo reconfiguration.

One appeal of active self-assembly models is that they allow formation of complex assemblies exponentially quickly by enabling insertion of new particles simultaneously throughout the assembly, a phenomenon observed in a wide range of biological systems [6, 19]. In contrast, passive self-assembly models such as the *abstract Tile Assembly Model (aTAM)* of Winfree [18] are limited to only polynomially fast growth [12].

* A preprint containing the omitted proof of Theorem 1 is available on arXiv: https://arxiv.org/abs/1411.0973.

[4] Not to be confused with *active* tile assembly models [8, 9, 11, 13, 16, 17] in which bond states change.

Of the active self-assembly models, both graph grammars and nubots are capable of a topologically rich set of assemblies and reconfigurations, but rely on stateful particles forming complex bond arrangements. In contrast, insertion systems consist of stateless particles forming a single chain of bonds. Indeed, all insertion systems are captured as a special case of nubots in which a linear polymer is assembled via parallel insertion-like reconfigurations, as in Theorem 5.1 of [19]. The simplicity of insertion systems makes their implementation in matter a more immediately attainable goal; Dabby and Chen [5, 6] describe experimental implementations in DNA.

1.1 Non-determinism in insertion systems

Previous work in insertion systems [6, 10] only considered deterministic systems, where each location accepts at most one monomer type. In the equivalent [10] model of context-free grammars, deterministic systems are those in which each non-terminal symbol appears on the left-hand side of a unique production rule. Allowing insertion sites in which different monomer types may be inserted gives rise to *non-deterministic* insertion systems. As a result of such sites, non-deterministic insertion systems may construct many distinct polymers.

In deterministic systems, the expected time of assembling length-n polymers is known [10] to have a tight lower bound of $\Omega(\log^{5/3}(n))$. In contrast, a simple non-deterministic system of just two monomer types is easily shown (see Section 3) to reduce expected assembly time to $\Theta(\log n)$. The cost of this reduced assembly time is a loss of precision: the two-monomer system yields infinitely many polymers of arbitrarily large lengths (i.e. is "pumpable" in grammar terminology), making it useless for targeted construction of polymers of a given length.

1.2 Our results

We consider whether non-determinism may be used to reduce assembly time without resulting in the complete loss of precision in the lengths of constructed polymers (i.e. without becoming "pumpable"). We answer this question in the affirmative, giving a non-deterministic insertion system that assembles length-n polymers in $\Theta(\log^{3/2}(n))$ expected time (Theorem 1) and a proof that this time is optimal systems that assemble a finite number of polymers (Theorem 2). However, the system still suffers from reduced precision: polymers of $\Theta(n)$ distinct lengths up to n are constructed (see Section 5 for further discussion).

The proof of Theorem 2 also implies a monomer type and time tradeoff for systems constructing a finite set of polymers: constructing a length-n polymer using k monomer types takes $\Omega(\log^2(n)/\sqrt{k})$ expected time (Lemma 4). This lemma implies that both our upper bound construction using $k = O(\log(n))$ monomer types (and assembling in $O(\log^{3/2}(n)$ expected time) and the upper bound construction of [10] using $k = O(\log^{2/3}(n))$ monomer types (and assembling in $O(\log^{3/2}(n)$ expected time) are optimal constructions at both ends of this tradeoff curve.

2 Definitions

An *insertion system* in the active self-assembly model of Dabby and Chen [6] carries out the construction of a linear *polymer* consisting of constant length *monomers*. A polymer grows incrementally by the insertion of a monomer at an *insertion site* between two existing monomers in the polymer, according to complementary bonding sites between the monomer and the insertion site.

An insertion system $\mathcal{S}$ is defined as a 4-tuple $\mathcal{S} = (\Sigma, \Delta, Q, R)$. The first element, Σ, is a set of symbols. Each symbol $s \in \Sigma$ has a *complement* s^*. We denote the complement of a symbol s as $\overline{s}$, i.e. $\overline{s} = s^*$ and $\overline{s^*} = s$. The set Δ is a set of *monomer types*, each assigned a *concentration*. Each monomer is specified by a quadruple $(a, b, c, d)^+$ or $(a, b, c, d)^-$, where $a, b, c, d \in \Sigma \cup \{s^* : s \in \Sigma\}$, and each concentration is a real number between 0 and 1. The sum of all concentrations in Δ must be at most 1. The two symbols $Q = (a, b)$ and $R = (c, d)$ are special two-symbol monomers that together form the *initiator* of $\mathcal{S}$. It is required that either $\overline{a} = d$ or $\overline{b} = c$. The *size* of $\mathcal{S}$ is $|\Delta|$, the number of monomer types in $\mathcal{S}$.

A *polymer* is a sequence of monomers $Qm_1m_2 \ldots m_nR$ where $m_i \in \Delta$ such that for each pair of adjacent monomers $(w, x, a, b)(c, d, y, z)$, either $\overline{a} = d$ or $\overline{b} = c$. The *length* of a polymer is the number of monomers, including Q and R, it contains. Each pair of adjacent monomer ends $(a, b)(c, d)$ form an *insertion site*. Monomers can be inserted into an insertion site $(a, b)(c, d)$ according to the following rules (see Figure 1):

1. If $\overline{a} = d$, then any monomer $(\overline{b}, e, f, \overline{c})^+$ can be inserted.
2. If $\overline{b} = c$, then any monomer $(e, \overline{a}, \overline{d}, f)^-$ can be inserted.[5]

A monomer is inserted after time t, where t is an exponential random variable with rate equal to the concentration of the monomer type. The set of all polymers *constructed* by an insertion system is recursively defined as any polymer constructed by inserting a monomer into a polymer constructed by the system, beginning with the initiator. Note that the insertion rules guarantee by induction that for every insertion site $(a, b)(c, d)$, either $\overline{a} = d$ or $\overline{b} = c$.

We say that a polymer is *terminal* if no monomer can be inserted into any insertion site in the polymer, and that an insertion system is *deterministic* if every polymer P constructed by the system is either P or is non-terminal and has length less than that of P (i.e. can become P).

3 Upper Bound for Assembly Time

It is natural to ask whether faster construction of polymers is possible in non-deterministic systems: systems that do not construct a single terminal polymer. A two-monomer-type insertion system consisting of the initiator $(s_1, s_2)(s_2^*, s_1^*)$

[5] In [6], this rule is described as a monomer $(\overline{d}, f, e, \overline{a})^-$ that is inserted into the polymer as $(e, \overline{a}, \overline{d}, f)$.

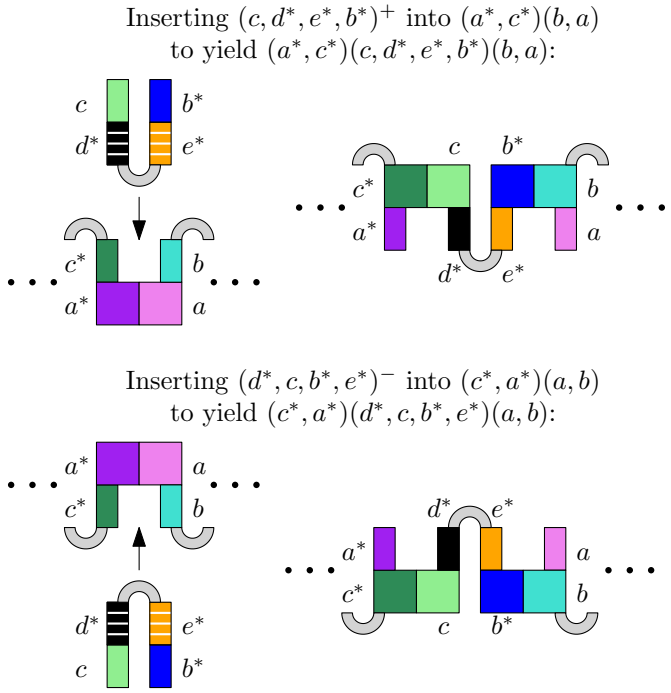

Fig. 1. A pictorial interpretation of the two insertion rules for monomers. Loosely based on Figure 2 and corresponding DNA-based implementation of [6].

and monomer types $(s_2^*, s_1^*, s_1, s_2)^+$, and $(s_2^*, x, x, s_2)^+$ simultaneously constructs polymers of all lengths $n \geq 3$ in expected time $O(\log n)$ via balanced insertion sequences of logarithmic length. Moreover, any polymer in any system has $\Omega(\log n)$ expected construction time, since every insertion takes $\Omega(1)$ expected time, and constructing a polymer of length n requires an insertion sequence of length at least $\lfloor \log_2(n-2) \rfloor$. So if assembling anything is permitted, then this two-monomer-type system is asymptotically optimal.

This leads to the question considered in this paper: to what extent is assembly time reduction possible in systems that only construct finite number of polymers? Our next result proves that even this relaxation is sufficient to improve assembly. The key idea of the construction is allow large sets of monomer types to "compete" to insert into a common insertion site first. This competition increases the total concentration of insertable monomer types, reducing the expected insertion time, but results in a non-deterministic system.

Theorem 1. *For any positive, odd integer r, there exists an insertion system constructing a finite set of polymers with $O(r^2)$ monomer types that constructs a polymer of length $n = 2^{\Theta(r^2)}$ in $O(\log^{3/2}(n))$ expected time. Moreover, the expected time has an exponentially decaying tail probability.*

Due to space constraints, the proof of Theorem 1 is omitted; we give a brief sketch here. The construction uses insertion sites to store the (two) variable values of a double for-loop, with short insertion sequences used to either increment the inner variable or increment the outer variable and reset the inner variable to 0.

Non-determinism is used to speed up loop increments in the following way: loops with different variable values "share" insertable monomer types, increasing the concentration of insertable monomer types for all insertion sites to decrease the expected time for the increment to take place. This comes with the tradeoff that not all monomer types lead to successful incrementing; some instead cause the loop to "break", leading to assembly of additional polymers of length less than n.

4 Lower Bound for Assembly Time

Here we show that the construction in the previous section is the optimal in expected assembly time and, for the given assembly time, optimal in the number of monomer types used (Theorem 2). A collection of intervening lemmas are used to prove bounds on the number of monomer types and expected time to carry out an *insertion sequence*: a sequence of monomer insertions where each insertion is into a site created by the previous insertion.

Observe that if two monomer types of the same sign are insertable into a common site, then the set of sites each can be inserted into is equal. Nearly all of the lemmas involve consideration of not only monomer types, but *insertion sets*: maximal sets of same-signed monomer types sharing a common set of insertion sites each can be inserted into.

The first several lemmas of the section are used to prove Lemma 4, a lemma describing the trade-off between the number of monomer types and expected construction time for systems constructing finite polymer sets. This lemma is combined with extremal bounds on the minimum number of monomer types and insertion sets to prove the final result.

Lemma 1. *Any insertion sequence of length l with no repeated insertion sites has $\Theta(l)$ sites of the form $(a, b)(\overline{c}, \overline{a})$ with $b \neq c$.*

Proof. Insertion sites have one of three forms:

Positive: $(a, b)(\overline{c}, \overline{a})$ with $b \neq c$.

Mixed: $(a, b)(\overline{b}, \overline{a})$.

Negative: $(b, a)(\overline{a}, \overline{c})$ with $b \neq c$.

We prove that every sequence of four consecutive insertion sites has at least one positive site. Consider such a sequence of four sites (and the three intervening insertions). If the first site is positive, we're done. If the first site is mixed, then the first monomer type inserted may be negative or positive. If a negative monomer type is inserted:

$$(a,b) \diamond (\overline{b},\overline{a})$$

$$(a,b)(\overline{c},\overline{a},a,d)(\overline{b},\overline{a})$$

Since the sequence does not repeat sites, $b \neq c,d$ and either second site, $(a,b)(\overline{c},\overline{a})$ or $(a,d)(\overline{b},\overline{a})$, is positive. If a positive monomer type is inserted:

$$(a,b) \diamond (\overline{b},\overline{a})$$

$$(a,b)(\overline{b},\overline{c},d,b)(\overline{b},\overline{a})$$

As before, $a,d \neq c$ since sites cannot repeat. So the next insertion must use a negative monomer type. We assume the left site is used next in the sequence (a symmetric argument works if the right site is used instead). The entire insertion sequence has the form:

$$(a,b) \diamond (\overline{b},\overline{a})$$

$$(a,b) \diamond (\overline{b},\overline{c},d,b)(\overline{b},\overline{a})$$

$$(a,b) \diamond (\overline{b},\overline{c})$$

$$(a,b)(\overline{e},\overline{a},c,f)(\overline{b},\overline{c})$$

As before, $e,f \neq b$ since sites cannot repeat. So the next site, either $(a,b)(\overline{e},\overline{a})$ with $b \neq e$ or $(c,f)(\overline{b},\overline{c})$ with $f \neq b$ is positive. So the third site in the sequence is positive. Finally, if the initial site is negative then the first monomer type inserted is negative:

$$(b,a) \diamond (\overline{a},\overline{c})$$

$$(b,a) \diamond (\overline{d},\overline{b},c,e)(\overline{a},\overline{c})$$

$$(b,a) \diamond (\overline{d},\overline{b})$$

We assume that the left site is used next in the sequence (a symmetric argument works if the right side is used instead). If $a \neq d$ then the second site is positive. Otherwise the second site is mixed, and by previous argument, at most two more insertions (a total of three) will take place until a positive site appears. So in the entire sequence of length l, a positive site appears at least once in every sequence of four consecutive sites. □

Lemma 2. *Any insertion sequence with no repeated insertion sites using k monomer types forming m insertion sets has length $O(m\sqrt{k})$.*

Proof. Let $\mathcal{S} = (\Sigma, \Delta, Q, R)$ be the insertion system containing the sequence. Relabel the symbols in $\Sigma \cup \{s^* : s \in \Sigma\}$ as $s_1, s_2, \ldots, s_{4k}$, with some of these symbols possibly unused. Note that this implies that for every s_i, $\overline{s_i} = s_j$ for some $j \in 1, 2 \ldots, 4k$. Let l be the length of the sequence. By Lemma 1, $\Theta(l)$ sites are *positive*: they have the form $(s_a, s_b)(\overline{s_c}, \overline{s_a})$ with $b \neq c$.

A bound of $\sum_{i=1}^{4k} \min(|\mathrm{L_i}|, |\mathrm{R_i}|) \leq 3m$. Let L_i and R_i be the sets of monomer types of the forms $(\_, \_, s_i, \_)^{\pm}$ and $(\_, \overline{s_i}, \_, \_)^{\pm}$, respectively, used in the insertion sequence. Each positive site $(s_i, s_b)(\overline{s_c}, \overline{s_i})$ consists of a left monomer in L_i and right monomer in R_i. Every occurrence of a positive site in the sequence is followed by the use of the left or right resulting site, e.g.:

$$(s_i, s_b) \diamond (\overline{s_c}, \overline{s_i})$$
$$(s_i, s_b) \diamond (\overline{s_b}, \overline{s_d}, s_e, s_c)(\overline{s_c}, \overline{s_i})$$
$$(s_i, s_b) \diamond (\overline{s_b}, \overline{s_d})$$

It is the case that d is unique for c, i.e. no two insertions into positive sites using the left resulting sites both use monomers of the form $(\overline{s_b}, \overline{s_d}, \_, \_)^{+}$, since such a pair of monomers implies the sequence repeats the insertion site $(s_a, s_b)(\overline{s_b}, \overline{s_d})$. A similar claim holds for e and b in the case that the right site is used. So inserting into the resulting site requires a monomer from a distinct insertion set $\{(\_, \overline{s_i}, s_d, \_)^{-} \in \Delta\}$ or, in the special case that $i = d$, $\{(\overline{s_b}, \_, \_, s_b)^{+} \in \Delta\}$.

The resulting sites require monomers from a number of distinct insertion sets equal to the sum of two values. First, the number of times the left side is used with a distinct c and a monomer is inserted into a site $(s_i, s_b)(\overline{s_b}, \overline{s_d})$ with d unique for c. Second, the number of times the right side is used with a distinct b and a monomer is inserted into a site $(s_e, s_c)(\overline{s_c}, \overline{s_i})$ with e unique for b. An assignment of left and right side usage that minimizes the number of distinct insertion sets needed is nearly equivalent to a minimum vertex covering of the following bipartite graph:

- A node $L_{(i,b)}$ for every site $(s_i, s_b)(\overline{s_c}, \overline{s_i})$ in the insertion sequence.
- A node $R_{(c,i)}$ for every site $(s_i, s_b)(\overline{s_c}, \overline{s_i})$ in the insertion sequence.
- An edge $(L_{(i,b)}, R_{(c,i)})$ for every site $(s_i, s_b)(\overline{s_c}, \overline{s_i})$ in the insertion sequence.

Selecting a vertex to cover a given edge corresponds to using the resulting left or right site of the edge's site, e.g. selecting $R_{(c,i)}$ to cover the edge $(L_{(i,b)}, R_{(c,i)})$ corresponds to using the resulting left site and inserting a monomer type of the form $(\_, \overline{s_i}, s_d, \_)^{-}$, where d is unique for c. By König's theorem (see [7, 1]), since the graph is bipartite, the size of a minimum vertex covering is equal to the size of a maximum matching, which is bounded from above by $\sum_{i=1}^{4k} \min(|\mathrm{L_i}|, |\mathrm{R_i}|)$.

However, an insertion set $\{(\_, \overline{s_e}, s_d, \_)^{-} \in \Delta\}$ corresponds to selecting both $R_{(c,i)}$, where d is unique for c, and $L_{(j,b)}$, where e is unique for b. So the number of insertion sets needed may be as little as half the size of the vertex cover of the bipartite graph. Additionally, one site may not be inserted into. So $\sum_{i=1}^{4k} \min(|\mathrm{L_i}|, |\mathrm{R_i}|) - 1 \leq 2\mathrm{m}$ and $\sum_{i=1}^{4k} \min(|\mathrm{L_i}|, |\mathrm{R_i}|) \leq 3\mathrm{m}$.

Maximizing insertion sequence length. Consider the number of positive sites y accepting some monomer type. We proved that $\Omega(l) = y$ and it is easily observed that $y \leq \sum_{i=1}^{4k} \min(\mathrm{m}, |\mathrm{L_i}| \cdot |\mathrm{R_i}|)$. We also proved that $\sum_{i=1}^{4k} \min(|\mathrm{L_i}|, |\mathrm{R_i}|) \leq 3m$ and it is easily observed that $\sum_{i=1}^{4k} \max(|\mathrm{L_i}|, |\mathrm{R_i}|) \leq$

$2k$, since each monomer type is in at most one L_i and one R_i. This gives the following set of constraints:

1. $\Omega(l) = \sum_{i=1}^{4k} \min(\mathrm{m}, |\mathrm{L_i}| \cdot |\mathrm{R_i}|)$.
2. $\sum_{i=1}^{4k} \min(|\mathrm{L_i}|, |\mathrm{R_i}|) \leq 3m$.
3. $\sum_{i=1}^{4k} \max(|\mathrm{L_i}|, |\mathrm{R_i}|) \leq 2k$.

Observe that $|L_i| \cdot |R_i| = \min(|\mathrm{L_i}|, |\mathrm{R_i}|) \cdot \max(|\mathrm{L_i}|, |\mathrm{R_i}|)$. Define two new variables $y_i = \min(|\mathrm{L_i}|, |\mathrm{R_i}|)$ and $z_i = \max(|\mathrm{L_i}|, |\mathrm{R_i}|)$ for an alternate formulation of the previous constraints:

1. $\Omega(l) = \sum_{i=1}^{4k} \min(\mathrm{m}, \mathrm{y_i z_i})$.
2. $\sum_{i=1}^{4k} y_i \leq 3m$.
3. $\sum_{i=1}^{4k} z_i \leq 2k$.

Relax y_i, z_i to be real-valued and let $W = \{i : y_i z_i > 0\}$. If $0 < y_i z_i, y_j z_j < m$ for some $i \neq j$ and $y_i = \max(\mathrm{y_i}, \mathrm{z_i}, \mathrm{y_j}, \mathrm{z_j})$, then $\min(\mathrm{m}, \mathrm{y_i z_i}) + \min(\mathrm{m}, \mathrm{y_j z_j}) < \min(\mathrm{m}, \mathrm{y_i}(\mathrm{z_i} + \varepsilon)) + \min(\mathrm{m}, \mathrm{y_j}(\mathrm{z_j} - \varepsilon))$ for sufficiently small $\varepsilon > 0$. More generally, if $0 < y_i z_i, y_j z_j < m$ for some $i \neq j$ then the values of y_i, z_i, y_j, z_j can be modified to increase $\sum_{i=1}^{4k} \min(\mathrm{m}, \mathrm{y_i z_i})$. Therefore the maximum value is achieved when $m = y_i z_i$ for all but at most one $i \in W$.

We claim that it cannot be that $y_i z_i = m$ for $6\sqrt{k}$ distinct values of i. By contradiction, assume so. So $|W| \geq 6\sqrt{k}$ and the average value of y_i for $i \in W$ must be less than $3m/(6\sqrt{k}) = m/(2\sqrt{k})$. So for a subset $W' \subseteq W$ with $|W'| \geq |W|/2 \geq 3\sqrt{k}$, $y_i \leq 2 \cdot m/(2\sqrt{k}) = m/\sqrt{k}$ for all $i \in W'$. For every $i \in W'$, because $y_i \leq m/\sqrt{k}$ and $y_i z_i = m$, it must be the case that $z_i \geq \sqrt{k}$. So $\sum_{i=1}^{4k} z_i \geq |W'| \cdot \sqrt{k} \geq 3k$, a contradiction with the constraint that $\sum_{i=1}^{4k} z_i \leq 2k$.

So the maximum value is achieved when $m = y_i z_i$ for all but at most one $i \in W$, with $|W| + 1 < 6\sqrt{k} + 1 < 7\sqrt{k}$. So $\sum_{i=1}^{4k} \min(\mathrm{m}, \mathrm{y_i z_i}) \leq (|W| + 1)m < 7m\sqrt{k}$. So $\Omega(l) = 7m\sqrt{k}$ and $l = O(m\sqrt{k})$. □

Lemma 3. *An insertion sequence of length l using monomer types from m insertion sets with no repeated insertion sites takes $\Omega(ml)$ expected time.*

Proof. By linearity of expectation, the total expected time of the insertions is equal to the sum of the expected time for each insertion. By Lemma 1, $\Theta(l)$ sites are both *positive*, i.e. they have the form $(s_a, s_b)(\overline{s_c}, \overline{s_a})$ with $b \neq c$, and accept the monomer types of a positive, non-empty insertion set.

Let m be the number of insertion sets formed by the monomer types inserted into these $\Omega(l)$ sites. Let $c_1, c_2, \ldots, c_m$ be the sums of the concentrations of the monomer types in these sets, and $x_1, x_2, \ldots, x_m$ be the number of times a monomer from each set is inserted in the subsequence. Then the total expected time for all of the insertions in the subsequence is $\sum_{i=1}^{m} x_i/c_i$. Moreover, these variables are subject to the following constraints:

1. $\sum_{i=1}^{m} x_i = \Omega(l)$ (total number of insertions is $\Omega(l)$).
2. $\sum_{i=1}^{m} c_i \leq 1$ (total concentration is at most 1).

Minimizing expected time. We now consider minimizing the total expected time subject to these constraints, starting with proving that $x_i/c_i = x_j/c_j$ for all $1 \leq i, j \leq m$. That is, that the ratio of the number of sites that accept an insertion set to the total concentrations of the monomer types in the set is equal for all sets. Assume, without loss of generality, that $x_i/c_i > x_j/c_j$ and $c_i, c_j > 0$. Then it can be shown algebraically that the following two statements hold:

1. If $c_j \geq c_i$, then for sufficiently small $\varepsilon > 0$, $\frac{x_i}{c_i} + \frac{x_j}{c_j} > \frac{x_i}{c_i+\varepsilon} + \frac{x_j}{c_j-\varepsilon}$.
2. If $c_j < c_i$, then for sufficiently small $\varepsilon > 0$, $\frac{x_i}{c_i} + \frac{x_j}{c_j} > \frac{x_i}{c_i-\varepsilon} + \frac{x_j}{c_j+\varepsilon}$.

Since the ratios of every pair of monomer types are equal,

$$\frac{c_i}{1} \leq \frac{c_i}{\sum_{i=1}^{m} c_i} = \frac{x_i}{\sum_{i=1}^{m} x_i} = O(x_i/l)$$

So $\Omega(l) = x_i/c_i$ and $\Omega(ml) = \sum_{i=1}^{m} x_i/c_i$. □

Lemma 4. *Any polymer of length n constructed by an insertion system with k monomer types constructing a finite set of polymers takes $\Omega(\log^2(n)/\sqrt{k})$ expected time.*

Proof. By Lemma 2, $n = 2^{O(m\sqrt{k})}$. So $m = \Omega(\log n/\sqrt{k})$. Constructing any polymer of length n requires an insertion system of length $l = \Omega(\log n)$. Then by Lemma 3, the expected time to construct any polymer of length n is $\Omega(ml) = \Omega(\log^2(n)/\sqrt{k})$.

Before proving the final result, we prove a helpful lemma showing that the number of insertion sets cannot be too much smaller than the number of monomer types:

Lemma 5. *Any insertion sequence of length l with no repeated insertion sites using k monomer types forming m insertion sets has $m = \Omega(\sqrt{k})$.*

Proof. Notice that this bound can only be obtained by assuming the monomer types are used to carry out an insertion sequence, since it is possible to have an arbitrarily large set of monomer types belonging to a single insertion set. The number of monomer types used is at most the length of the insertion sequence ($k \leq l$), and the remainder of the proof is spent proving that the number of insertion sites in a system with m insertion sets is $O(m^2)$ ($l = O(m^2)$), giving the desired inequality.

Let $\mathcal{S} = (\Sigma, \Delta, Q, R)$ be the insertion system containing the sequence. Relabel the symbols in $\Sigma \cup \{s^* : s \in \Sigma\}$ as $s_1, s_2, \ldots, s_{4k}$, with some of these symbols possibly unused. By Lemma 1, $\Omega(l)$ sites are *positive*: they have the form $(s_a, s_b)(s_c, \overline{s_a})$ with $b \neq c$.

Since the second monomer inserted to create the site must be negative, each positive site consists of at least one negative monomer type. Let L_i^- and R_i^- be the sets of monomer types of the forms $(\_, \_, s_i, \_)^-$ and $(\_, \overline{s_i}, \_, \_)^-$, respectively, used in the insertion sequence of length l. For a specific i, there exists a site of the form $(s_i, s_b)(s_c, \overline{s_i})$ only if $|L_i^-|+|R_i^-| > 0$. So the number of values of i such that a site of the form $(s_i, s_b)(s_c, \overline{s_i})$ exists is at most $\sum_{i=1}^{4k} |L_i^-|+\sum_{i=1}^{4k} |R_i^-|$. Since all monomer types of a negative insertion set belong to the same L_i^- and R_i^-, $\sum_{i=1}^{4k} |L_i^-| + \sum_{i=1}^{4k} |R_i^-| \leq 2m$.

Next, observe there are at most m sites of the form $(s_i, s_b)(s_c, \overline{s_i})$ that accept a monomer, since each site requires a monomer from a different positive insertion set. So the total number of positive sites that accept a monomer is at most $2m \cdot m = 2m^2$. Since there are $\Omega(l)$ positive sites in the insertion sequence, $\Omega(l) = 2m^2$ and $l = O(m^2)$. □

Theorem 2. *Any polymer of length n constructed by an insertion system constructing a finite set of polymers takes $\Omega(\log^{3/2}(n))$ expected construction time. Moreover, constructing a polymer of length n in $\Theta(\log^{3/2}(n))$ expected time requires using $\Omega(\log n)$ monomer types.*

Proof. First, observe that constructing a polymer of length n in a system constructing a finite set of polymers involves an insertion sequence of length $\log_2(n) \leq l$ with no repeated sites. By Lemmas 2 and 5, $\log_2(n) = O(m^2)$ and so $m = \Omega(\sqrt{\log n})$ Then by Lemma 3, carrying out the insertion sequence and completing the construction of the polymer takes $\Omega(ml) = \Omega(\log^{3/2}(n))$ expected time and by Lemma 4, $k = \Omega(\log n)$. □

5 Open Problems

The results of in this paper, combined with those of [10] describe the landscape of efficient polymer construction using insertion systems:

- Trivial systems of just a few polymers can construct polymers of arbitrary length in optimal time, but with the caveat that the growth is uncontrolled and the systems construct infinite set of polymers.
- Deterministic construction a polymer of length n requires $\Omega(\log^{2/3}(n))$ monomer types and $\Omega(\log^{5/3}(n))$ expected time, and both of these are achievable simultaneously.
- The intermediate situation of constructing finite sets of polymers is more intricate – polymers can be constructed faster, but with the trade-off of using more monomer types *and* non-determinism.

In our system achieving $O(\log^{3/2}(n))$ expected construction time (Theorem 1), an exponential number ($2^{\Theta(n \log \log n)}$) of "junk" terminal polymers are constructed. Since achieving such speed requires significantly fewer insertion sets than monomer types, some junk is necessary – but how much? One approach to proving a lower bound is to prove that insertion sites accepting large insertion

sets imply a large number of terminal polymers. We have been unable to prove such an implication even in the simplest case:

Conjecture 1. Every deterministic system with no unused monomer types has exclusively singleton insertion sets.

Since assembling a polymer in $o(\log^{5/3}(n))$ expected time requires that $\Omega(\log n)$ insertions along most insertion sequences are non-deterministic, the previous conjecture implies that any improvement in speed comes with an exponential number of junk terminal polymers:

Conjecture 2. Any system constructing a polymer of length n in $O(\log^{3/2}(n))$ expected time constructs a set of $2^{\Omega(n)}$ polymers.

Setting aside non-determinism, the trade-off between monomer types and construction time has a lower bound (Lemma 4) with matching upper bounds only at the extremes. Does there exist a parameterized system matching the lower bound across the entire range?

Conjecture 3. For every combination of n and k such that $\log_2^{2/3}(n) \leq k \leq \log_2(n)$, there exists a system with k monomer types that constructs a polymer of length n in $O(\log^2 n/\sqrt{k})$ time.

References

1. Bondy, J.A., Murty, U.S.R.: Graph Theory with Applications. Elsevier, New York (1976)
2. Chen, H.L., Doty, D., Holden, D., Thachuk, C., Woods, D., Yang, C.T.: Fast algorithmic self-assembly of simple shapes using random agitation. In: DNA Computing and Molecular Programming (DNA 2014), LNCS, vol. 8727, pp. 20–36. Springer (2014)
3. Chen, M., Xin, D., Woods, D.: Parallel computation using active self-assembly. vol. 14, pp. 225–250 (2015)
4. Chin, Y.R., Tsai, J.T., Chen, H.L.: A minimal requirement for self-assembly of lines in polylogarithmic time. In: DNA Computing and Molecular Programming (DNA 2017). LNCS, vol. 10467, pp. 139–154. Springer (2017)
5. Dabby, N.: Synthetic molecular machines for active self-assembly : prototype algorithms, designs, and experimental study. Ph.D. thesis, Caltech (2013)
6. Dabby, N., Chen, H.L.: Active self-assembly of simple units using an insertion primitive. In: Proceedings of 24th ACM-SIAM Symposium on Discrete Algorithms (SODA). pp. 1526–1536 (2013)
7. Diestel, R.: Graph Theory. Springer, Berlin (2005)
8. Gautam, V.K., Haddow, P.C., Kuiper, M.: Reliable self-assembly by self-triggered activation of enveloped DNA tiles. In: Dediu, A.H., Martín-Vide, C., Truthe, B., Vega-Rodríguez, M.A. (eds.) Theory and Practice of Natural Computing. LNCS, vol. 8273, pp. 68–79. Springer Berlin Heidelberg (2013)
9. Hendricks, J., Padilla, J.E., Patitz, M.J., Rogers, T.A.: Signal transmission across tile assemblies: 3D static tiles simulate active self-assembly by 2D signal-passing tiles. In: DNA Computing and Molecular Programming. LNCS, vol. 8141, pp. 90–104. Springer (2013)

10. Hescott, B., Malchik, C., Winslow, A.: Tight bounds for active self-assembly with an insertion primitive. Algorithmica 77(2), 537–554 (2017)
11. Jonoska, N., Karpenko, D.: Active tile self-assembly, part 1: universality at temperature 1. International Journal of Foundations of Computer Science 25(2), 141–163 (2014)
12. Keenan, A., Schweller, R., Sherman, M., Zhong, X.: Fast arithmetic in algorithmic self-assembly. Tech. rep., arXiv (2013)
13. Keenan, A., Schweller, R., Zhong, X.: Exponential replication of patterns in the signal tile assembly model. In: DNA Computing and Molecular Programming (DNA 2013). LNCS, vol. 8141, pp. 118–132. Springer (2013)
14. Klavins, E.: Universal self-replication using graph grammars. In: Proceedings of International Conference on MEMS, NANO, and Smart Systems. pp. 198–204 (2004)
15. Klavins, E., Ghrist, R., Lipsky, D.: Graph grammars for self assembling robotic systems. In: Proceedings of the International Conference on Robotics and Automation (ICRA). vol. 5, pp. 5293–5300 (2004)
16. Majumder, U., LaBean, T.H., Reif, J.H.: Activatable tiles: Compact, robust programmable assembly and other applications. In: DNA Computing and Molecular Programming (DNA 2008), LNCS, vol. 4848, pp. 15–25. Springer (2008)
17. Padilla, J.E., Liu, W., Seeman, N.C.: Hierarchical self assembly of patterns from the robinson tilings: DNA tile design in an enhanced tile assembly model. Natural Computing 11(2), 323–338 (2012)
18. Winfree, E.: Algorithmic Self-Assembly of DNA. Ph.D. thesis, Caltech (1998)
19. Woods, D., Chen, H.L., Goodfriend, S., Dabby, N., Winfree, E., Yin, P.: Active self-assembly of algorithmic shapes and patterns in polylogarithmic time. In: Proceedings of 4th Conference on Innovations in Theoretical Compuer Science (ITCS). pp. 353–354 (2013)