

Errors in DNA Self-Assembly By Synthesized Tile Sets

X. Ma, M. Hashempour, Y.B. Kim and F.Lombardi
Dept of Electrical and Computer Engineering
Northeastern University
Boston, MA 02115
{xma,masoud, ybk,lombardi}@ece.neu.edu

Abstract

This paper presents a study of errors that occur in DNA self-assembly using synthesized tile sets for template manufacturing. It is shown that due to the reduced size, aggregates assembled by a synthesized tile set are not error-free as those assembled by maximum-sized (referred to as a trivial tile set) as well as non-synthesized tile sets. Compared with non-synthesized tile sets, aggregates assembled using synthesized tile sets also have a higher error rate at high tile concentration, but they exhibit a lower rate at low tile concentration. In this paper, it is also shown that errors in synthesized tile sets tend to appear in clusters. This is very different from non-synthesized tile sets of non-maximum size in which growth errors are mostly random in nature. These findings are discussed and evaluated for nano manufacturing (such as for template self-assembly of a QCA circuit).

Index Terms: DNA Self-assembly, error analysis, nanoscale manufacturing, tile set, emerging technologies.

1 Introduction

For nano manufacturing, DNA self-assembly relies on defining a set of tiles and related features (such as bonds) to target a specific pattern as a template. Self-assembly must utilize simple tiles and possibly, a tile set of small cardinality to reduce the laboratory work required for implementation. Some patterns, such as the Sierpinski Triangle [1] and the binary counter [2] can be generated by a compact tile set. However, only a limited number of such patterns are available while, as a manufacturing method, the ability to assemble arbitrary finite-size patterns as templates is highly sought. Arbitrary patterns can be built by tile sets generated by assigning every tile in the aggregate a unique type. The design of tile sets of this type is intuitive; however, these tile sets have a large cardinality (i.e. the maximum number of tile types in the tile set). This is equal to the number of pixels in the pattern. Such tile set is referred to as a *trivial tile set*. A process referred to as *tile synthesis* has been proposed in [3] to simplify the trivial tile set and reduce its cardinality. Tile synthesis starts from a trivial tile set of a finite size pattern and generates a tile set for the same pattern with a reduced (and in some cases optimal) cardinality. Finite size patterns are suitable for nano-scale manufacturing as they are modular and allow the assembly of diverse periodic structures such as scaffolds and templates for molecular electronics as well as emerging technologies [4] [5]. Synthesized tile sets are therefore a necessity for a practical application of DNA self-assembly to nanoscale manufacturing.

Unfortunately, the application of DNA self-assembly is severely limited by its high error rates. It has been shown that the error rate in a DNA self-assembly process can be as high as 10% [6], usually involving a large number of DNA molecules arranged into tiles in non-synthesized sets. No error analysis has been reported in the technical literature on synthesized tile sets. This paper addresses the issues revolving around errors that occur in the assembly using synthesized tile sets

for template manufacturing. It is shown by simulation that a synthesized aggregate has an error rate higher than a non-synthesized aggregate at high tile concentration. However, its error rate becomes lower than a non-synthesized aggregate at low tile concentration. The causes behind the difference between synthesized and non-synthesized tile sets are then discussed. It is also observed that upon occurrence of an error in a pattern generated by a synthesized tile set, a cluster of erroneous tiles (sited on facets) appears. This scenario is different for non-synthesized tile sets (such as the tile sets from [1][2]) in which growth errors are mostly random in nature.

2 Review of DNA Self-Assembly

DNA by itself does not have desirable electrical properties. However, the programmability of DNA in the form of tiles can be exploited to facilitate the manufacturing of templates for circuits to deposit electronic devices, such as molecular transistors or quantum-dot cellular automata (QCA) cells [7][8].[9][10][11] have reported the use of DNA self-assembly to build nanostructures as scaffolds, templates and interconnects for molecular electronics. The basic principle is to utilize the programmability of DNA tiles to self-assemble into periodic and aperiodic lattice structures [9]. Then, it is possible for other molecular electronic devices or quantum dots to selectively attach through deposition to the lattice structure. An obvious advantage of this technique is that a massive parallel assembly is possible, such that millions of copies of the desired structure can be built simultaneously [9]. This technique has been advocated as a viable method for future “bottom-up” nanofabrication also applicable to emerging technologies such as QCA [4][5]. The *abstract Tile Assembly Model (aTAM)* provides the basis for analyzing self-assembly in ideal cases [16][17]. A *tile set* consists of a finite set of unique *tiles* that are used to self-assemble into a DNA *crystal*. A tile is assumed to be square and tiles can not rotate by assumption. Each of the four sides of a tile has a *bond type*. The bond types of a tile determine the uniqueness of the tile. Each bond type has an associated *bond strength*. Two bonds of the same type can glue (i.e. bond) together, with a corresponding strength. It is assumed that the strength between different bond types is always 0. aTAM can be extended to other tile structures (such as those having more than four sides) with no loss of correctness and generality.

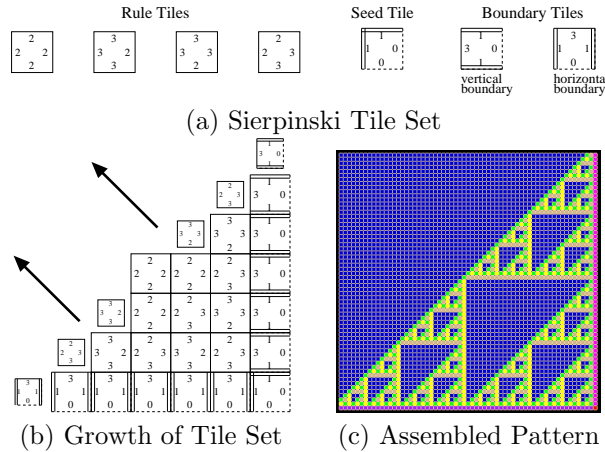


Figure 1. The Sierpinski Triangle Tile Set

In the ideal process assumed by the aTAM model, self-assembly always begins with a *seed tile*. A tile can be added to the existing crystal when its total bond strength to the crystal is greater than or equal to 2. An empty location in which the total bond strength of its adjacent sides is ≥ 2 is called a *growth site* because a new tile whose sides match the adjacent sides, can attach to this location in aTAM. The crystal generated from the seed tile via a series of legal (correct) tile

additions is referred to as the *produced assembly*. The commonly employed convention uses line(s) to indicate the strength of a bond type. The *integer* in the tile denotes the *bond type*. The example in Figure 1 shows a widely studied tile set for nanomanufacturing. The presentation of the tile set in Figure 1(a) follows the convention introduced above. The assembly process of this tile set according to aTAM is given in Figure 1(b) and the final assembly as shown in Figure 1(c) is a pattern generally referred to as the *Sierpinski triangle*.

While for the ideal scenario, aTAM is adequate for analysis and design of tile sets, a new model is required for dealing with related issues (such as growth speed and error analysis). In a more realistic assembly process, *insufficient attachments* and *tile dis-association* can occur. An *insufficient attachment* (or mismatch) is an event in which a tile attaches to the crystal with total bond strength less than 2. In tile dis-association, a tile falls off from a crystal being activated by thermal noise. The *kinetic Tile Assembly Model (kTAM)* provides a framework to analyze and simulate this non-ideal scenario of self-assembly. The kTAM model includes rates for both *association and dis-association* of tiles from the crystal. In this model, the on-rate (association) r_{on} is determined only by the tile concentration, that is represented by the parameter G_{mc} . The off-rate r_{off} (dis-association) is determined by the total bond strength and the parameter G_{se} for the temperature. A simulation tool *Xgrow* has been developed in [17]; Xgrow implements the kTAM model to emulate the self-assembly of a DNA assembled crystal. Different types of errors can be observed in DNA self-assembly [6]. The growth error refers to the type of error in which a *weakly-binding* tile (with total bond strength less than 2) attaches to a site where another strongly-binding tile (total bond strength ≥ 2) should have been added. The *facet roughening error* refers to the error such that a weakly-binding tile attaches to a site in which no tile should yet be added [6]. This will cause an uncontrolled spontaneous growth on a facet.

For nano manufacturing, DNA self-assembly relies on defining a set of tiles and related features (such as bonds) to target a specific finite-size pattern. Self-assembly must utilize simple tiles and possibly, a tile set of small cardinality to reduce any laboratory work required for implementation. To generate a pattern by self-assembly, a tile set can be computed by considering the combinatorial arrangements of the tiles and bond types in the aggregate; however, such set is trivial (i.e. large and intuitive by construction), hence a process amenable to simplification and set reduction by computation, is often required. Such process is generally referred to as *tile synthesis* [3]. In [3], issues revolving around the synthesis of tile sets for DNA self-assembly were addressed. Tile synthesis is an optimization problem targeting specific features such as the number of tiles and/or bonds. It has been defined as a combinatorial optimization problem referred to as the Pattern self-Assembly Tile-set Synthesis (PATS) problem. A graph model has been proposed for analyzing the tile sets assembling a specified pattern. The PATS problem is equivalent to the minimum graph coloring problem, hence in the general case it remains NP complete. Two greedy algorithms (referred to as PATS_Tile and PATS_Bond) have been proposed for synthesis of tile sets for self-assembly.

3 Errors in Synthesized Tile Sets

In DNA self-assembly, different types of error have been observed in simulated and laboratory trials; moreover, errors occur in kTAM at a rate between 1% and 10%. Errors impose limitations on the application of DNA self-assembly. This paper focuses on the implications of errors in synthesized tile sets for self-assembly. As the tiles in the trivial set have unique bonds on each side, an insufficient attachment will not necessarily result in an erroneous assembly. In a synthesized tile set, the bond types are no longer unique, so errors may appear in the final assembly. The tile sets of some patterns are available, hence referred in this paper to as non-synthesized test sets. However, self-assembly from non-synthesized patterns may also suffer from errors because bond types are not unique to every tile.

While the above analysis is intuitive, an investigation on errors in self-assembly using different types of tile sets is pursued using Xgrow as simulation tool. In this study, the simulation parameters are set as follows: $G_{se} = 10$, G_{mc} varies from 18.0 to 19.5. The simulation window has been set to

128 × 128 growth sites for the tiles and the simulation time is set to 80,000 seconds.

Tile sets for 70 different random patterns of various pattern sizes [3] have been simulated. The tile synthesis process results on average in a decrease of 75.6% in bond types and 43.5% in tile types, compared to trivial tile sets. Each tile set was simulated 30 times. Simulation results of all synthesized tile sets under the same G_{mc} are used to calculate the average error rate. The average error rate is plotted against G_{mc} in Figure 2. Simulation results of all trivial tile sets are used to calculate the average error rate under different G_{mc} . The error rate of a trivial tile set is at least 2 orders of magnitude smaller than in the corresponding synthesized tile set.

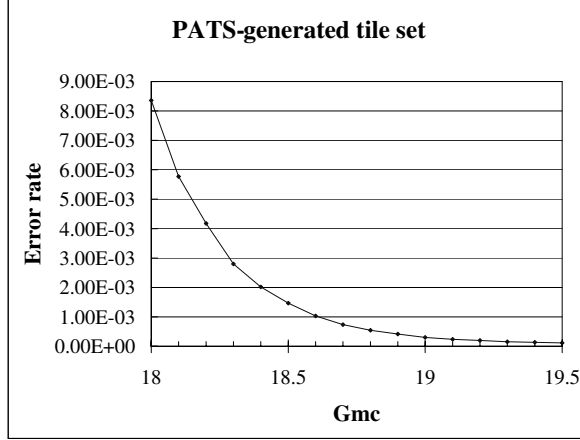
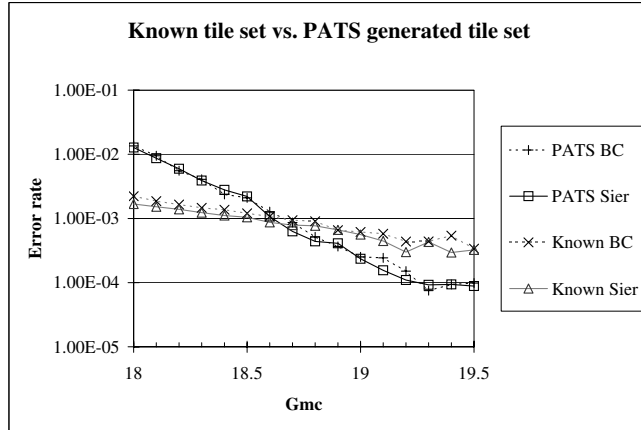


Figure 2. Error rate of synthesized tile sets for random patterns

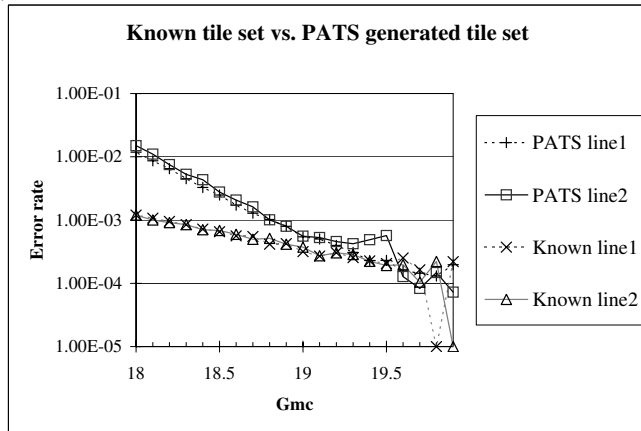
Next, the error rates of the synthesized tile sets for the Sierpinski triangle and the binary counter patterns [3] are compared with the non-synthesized tile sets for the same patterns (given in [14]). As shown in Figure 3(a), the two tile sets generated by PATS (*PATS BC* and *PATS Sier*) have a similar error rate and the two non-synthesized tile sets (*Known BC* and *Known Sier*) have similar error rates too. The PATS-generated tile sets have a higher error rate than the non-synthesized tile sets under small values of G_{mc} . However, when G_{mc} is large (corresponding to the scenario of growth at low error rate), a synthesized tile set has an error rate lower than the non-synthesized tile set of the same pattern. Figure 3(b) shows the comparison of PATS generated tile sets and non-synthesized tile sets for two other patterns (line1 and line2). These are patterns of periodic vertical lines with controlled spacing (as required by templates for self-assembly of nano-interconnects). As plotted in Figure 3(b) similar results are observed; further analysis of the error rate is pursued in the next section to understand its behavior.

Assembled templates made of replicating the basic pattern are also studied. Assembled patterns of the Sierpinski triangle are shown in Figure 4. Figure 4(a) shows the assembly from the synthesized 12 × 12-pixel Sierpinski tile set, with mismatched tiles highlighted in Figure 4(b). The simulation results of the non-synthesized tile set for the Sierpinski triangles are given in Figure 4(c)-(d). It can be observed that errors in the PATS-generated tile set appear in clusters, while errors are mostly dispersed by using a non-synthesized tile set.

Simulation has also been conducted using a tile set synthesized for manufacturing the layout of a QCA (Quantum-dot Cellular Automata) [4] circuit. Recently [5], self-assembly has been advocated as a technique for building templates on which QCA cells can be attached for nano circuit manufacturing. Figure 5 shows the pattern for replicating the co-planar crossing circuit generated using a synthesized tile set with $G_{mc} = 18.2$ and $G_{se} = 10$. Clustering of errors can also be observed in this pattern as example of a template for nanomanufacturing. The observed error clustering in synthesized tile sets confirms the findings established in previous experiments.



(a) Tile sets for patterns “Sierpinski” and “Binary Counter”



(b) Tile sets for patterns “line1” and “line2”

Figure 3. Error rate of tile sets for various self-assembly patterns

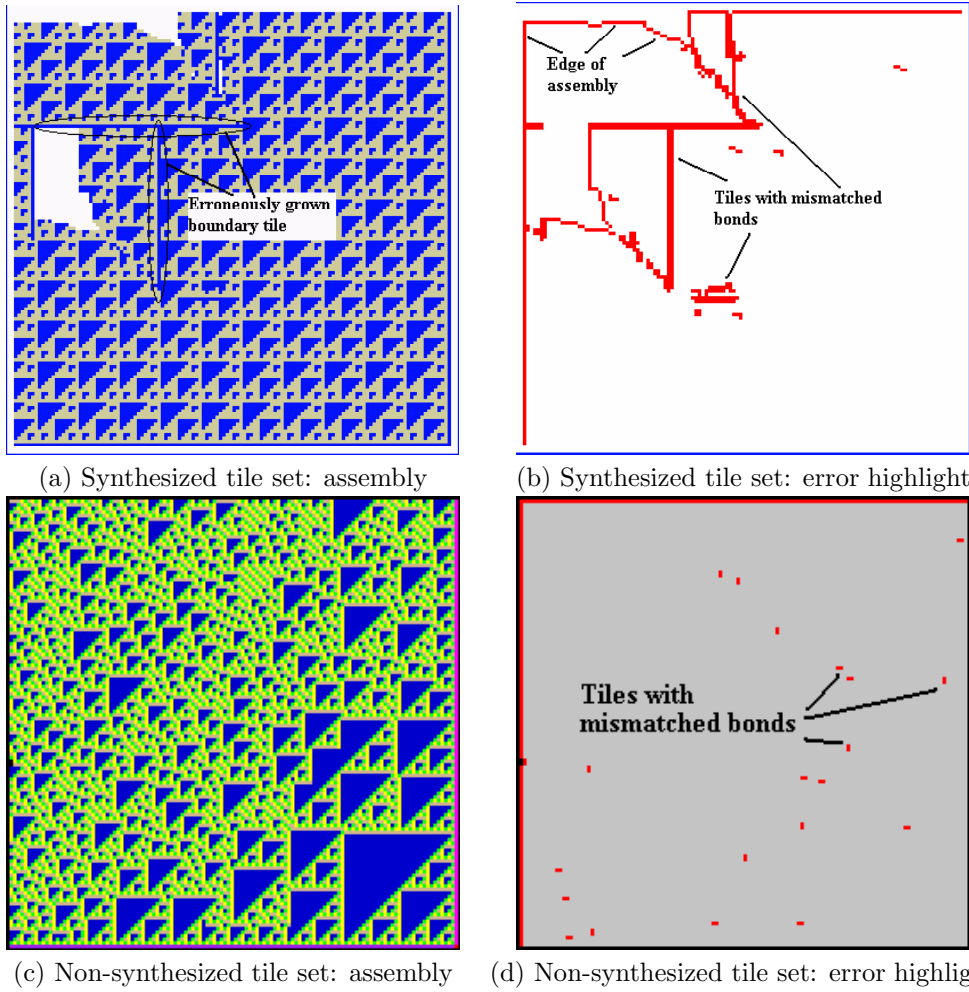


Figure 4. Examples for comparing the error occurrence in self-assembly of synthesized and non-synthesized tile sets

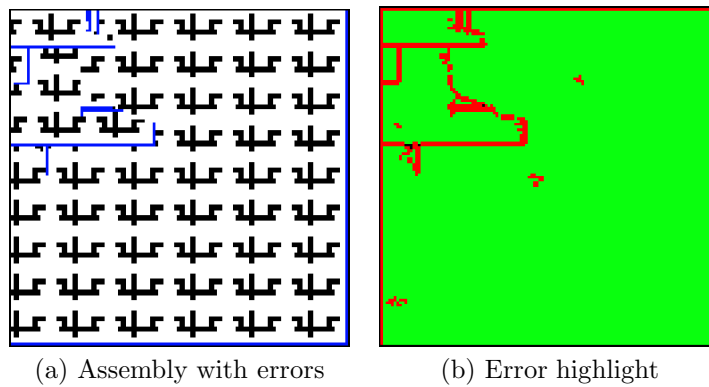


Figure 5. Error occurrence in self-assembly of synthesized tile set for QCA coplanar crossing layout pattern

Name of Finite Pattern	# of Bonds	# of Tiles	Probability of Matching
AB_15	70	164	0.1592
AB_6	18	36	0.1921
BinaryCounter_15	73	169	0.1919
BinaryCounter_6	17	35	0.3176
line1_15	74	151	0.1784
line1_6	8	15	0.1422
line2_15	74	166	0.2341
line2_6	22	36	0.2870
Sierpinski_12	53	116	0.1544
Sierpinski_6	19	35	0.2718
Sierpinski_8	25	57	0.3192

Table 1. Probability of having a matching tile type for a random growth site

4 Discussion

Initially, consider that self-assembly from a trivial tile set is virtually error-free compared to those from a synthesized tile set and a non-synthesized tile set. Let's define a *matching growth site* as a growth site that can perfectly match one of the tile types in the tile set. If the aggregate grows with no error, then every growth site is a matching site. However, after the occurrence of an error, the new growth sites have neighbouring conditions different from those encountered in the error-free growth. For a PATS-generated tile set, there is a high probability that a new growth site is not a matching growth site. When an unmatching growth site appears, additional insufficient attachments may occur in the DNA self-assembly process. an unmatching growth site appears not only at the locations next to the initial error, but any growth site that has a (bond) dependence with the initial erroneous tile, can be affected. Assume that the bond types that appear on the south and east sides of the error-affected growth sites, are random (as in the worst case scenario). So, an error-affected growth but matching site has the same probability as a randomly generated bond type combination (made of 2 bond types) for matching to the east and south sides of some tiles in the tile set. Let the presence of bond types have the same probability distribution on the east/west and south/north sides (as in constant monomer concentrations) and the total number of bond types be denoted by A . Also on average, a bond type appears on the south sides of r different tile types (as an approximation, a bond type also appears on the east/north/west sides of r different tile types); moreover, the probability of an error-affected growth matching site is $\frac{r}{A}$. For example, if a bond type appears on the south side of r different tile types in the tile set, a growth site is matching only if its east side matches one of the r tile types. By assumption, the bond type on the east side of the growth site is random (i.e., it can be any of the A types), so the probability of having a matching growth site is $\frac{r}{A}$. This assumption provides a worst case scenario by which errors can be considered. Eleven synthesized tiles sets that have been generated by PATS for the *Sierpinski*, the *AB grid*, the *barcode* and the *binary counter* patterns, have been investigated (the numeric suffix identifies the dimension of the square pixel area as a finite size pattern in the template) [3]. Table 1 shows the probability of having a growth site to match any available tile type, provided each tile located south and east to the site is selected randomly from the tile set (in the ideal case of no error). The probability of having a matching growth site is in the range of 15% to 30%. Differently from the synthesized tile sets, non-synthesized tile sets are designed such that any new growth site is a matching growth site. So, the probability of having a matching growth site is 100%; this probability is a significant measure that affects the correct growth of the self-assembly.

Next, consider that errors cluster in assemblies from synthesized tile sets. Figure 6 shows part of an assembly using a tile set synthesized for the 6×6 pixel area of the Sierpinski pattern. This example illustrates that unmatching growth sites in an error affected area result in clustered errors.

Figure 6(a) shows the tiles in an error-free assembly. When an initial insufficient attachment occurs (as in Figure 6(b)), this results in some growth sites for which it is not possible to find compatible tile types. Continued growth of the self-assembly must include more mismatched tiles in these incompatible growth sites. Figure 6(c) shows an assembly that can occur as outcome of the initial error. The mismatched error tiles are *clustered*. Hence, an *error-affected area* is generated; this area is located at the north-west of the initial error (according to the assumed direction of growth). By comparison, the non-synthesized tile sets have 100% probability of having a matching growth site. So an initial mismatch does not result in more mismatched tile attachments. Clustering can affect error generation in two ways; it decreases the rate of error generation due to the lower probability of attaching a series of mismatches. This effect is similar to the error tolerance mechanism of proofreading DNA tile sets [6]. Also, if an initial error stays in the assembly, it results in a large number of additional errors. This explains the difference in error rates between synthesized tile sets and non-synthesized tile sets under different values of G_{mc} .

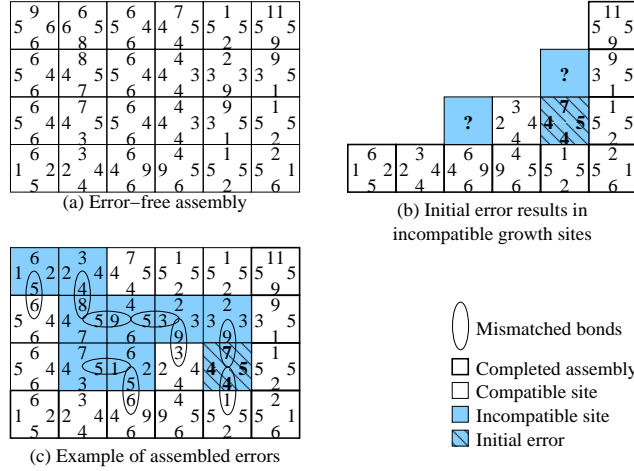


Figure 6. An initial error in the synthesized tile set assembly results in a cluster of errors

5 Conclusion

This paper has dealt with DNA self-assembly using different tile sets (such as synthesized, trivial and non-synthesized tile sets) and the implications of set cardinality on error generation. While generating an almost error free aggregate, the large size of trivial tile sets make them impractical, hence the need for synthesis. Compared with non-synthesized tile sets, patterns grown by synthesized tile sets have a higher error rate at low G_{mc} , but a lower error rate at large G_{mc} . This paper has also shown that errors in an aggregate assembled from a synthesized tile set, are mostly clustered, i.e. when an initial mismatched tile occurs in the assembly, it causes a series of growth errors in the area located to the north-west of the initial mismatched tile (this area is referred to as the error-affected area). When G_{mc} is small, error clustering causes an increase in the number of mismatched tiles in the self-assembly. When G_{mc} is large, growth of erroneous tiles is reduced by error clustering due to the lower probability of generating a series of errors.

Further work is needed to analyze the errors observed in these experiments. A probability based error model is being considered to characterize and predict errors in DNA self-assembly from synthesized tile sets. Also, research is being pursued to fully characterize the implications of errors in DNA self-assembly on the yield for manufacturing QCA circuits.

References

- [1] P. W. K. Rothmund, N. Papadakis, and E. Winfree, “Algorithmic self-assembly of dna sierpin-ski triangles,” *PLoS Biology*, vol. 2782, no. 12, p. e424, 2004, doi:10.1371/journal.pbio.0020424.
- [2] R. D. Barish, P. W. K. Rothmund, and E. Winfree, “Two computational primitives for algorithmic self-assembly: Copying and counting,” *Nano Letters*, vol. 5, no. 12, pp. 2586–2592, 2005.
- [3] X. Ma and F. Lombardi, “Synthesis of tile sets for DNA self-assembly,” *IEEE Tran. on CAD of ICAD*, vol. 27, no. 5, pp. 963–967, 2008.
- [4] J. Huang and F. Lombardi, “Design and test of digital circuits by quantum-dot cellular automata,,” *Artech House*, 2007.
- [5] H. H. Crocker M., Niemier M. and L. M., “Molecular qca design with chemically reasonable constraints,” *ACM JETC*, vol. 4, no. 2, 2008.
- [6] E. Winfree and R. Bekbolatov, “Proofreading tile sets: Error correction for algorithmic self-assembly,” in *Proc. 9th Intl. Workshop on DNA Computing*, 2003, pp. 108–126.
- [7] W. Hu, K. Sarveswaran, L. M., and G. Bernstein, “High-resolution electron beam lithography and DNA nano-patterning for molecular QCA,” *IEEE Tran. on Nanotechnology*, vol. 5, no. 3, pp. 312–316, 2005.
- [8] G. Bernstein, W. Hu, Q. Hang, K. Sarveswaran, and L. M., “Electron beam lithography and liftoff of molecules and DNA rafts,” in *Proc. IEEE conf. on Nanotechnology*, 2004, pp. 201–203.
- [9] S.-H. Park, H. Yan, J. Reif, T. LaBean, and G. Finkelstein, “Electronic nanostructures templated on self-assembled DNA scaffolds,” *Nanotechnology*, vol. 15, pp. S525–S527, 2004.
- [10] S.-H. Park, P. Yin, Y. Liu, J. Reif, T. LaBean, and H. Yan, “Programmable DNA self-assemblies for nanoscale organization of ligands and proteins,” *Nano Letters*, vol. 5, no. 4, pp. 729–733, 2005.
- [11] S.-H. Park, R. Barish, H. Li, J. Reif, G. Finkelstein, H. Yan, and T. LaBean, “Three-helix bundle DNA tiles self-assemble into 2D lattice or 1D templates for silver nanowires,” *Nano Letters*, vol. 5, no. 4, pp. 693–696, 2005.
- [12] L. Adleman, “Molecular computation of solutions to combinatorial problems,” *Science*, vol. 266, pp. 1021–1024, 1994.
- [13] H. Wang, “An unsolvable problem on dominoes,” Harvard Computation Laboratory, Technical Report BL30(II-15), 1962.
- [14] E. Winfree, “DNA computing by self-assembly,” *NAE’s The Bridge*, vol. 33, no. 4, pp. 31–38, 2003.
- [15] M. Lagoudakis and T. LaBean, “2d DNA self-assembly for satisfiability,” in *Proc. 5th DIMACS Workshop on DNA Based Computers*, ser. DIMACS Series in Discrete Mathematics and Theoretical Computer Science, E. Winfree and D. Gifford, Eds., 1999, pp. 141–154.
- [16] E. Winfree, “Self-healing tile sets,” in *Nanotechnology: Science and Computation*. Springer, 2006, pp. 55–78.
- [17] —, “Xgrow homepage,” Available online: www.dna.caltech.edu/Xgrow/.