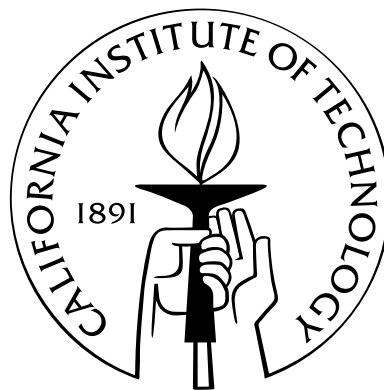


The Self-Replication and Evolution of DNA Crystals

Thesis by
Rebecca Schulman

In Partial Fulfillment of the Requirements
for the Degree of
Doctor of Philosophy



California Institute of Technology
Pasadena, California

2007
(Defended May 11, 2007)

© 2007
Rebecca Schulman
All Rights Reserved

Acknowledgements

I came to Caltech a scatterbrained but enthusiastic young scientist. Without the constant nurturing and tutelage of my PhD advisor, Erik Winfree, I can't imagine what would have happened. Erik's gifts are many – a generous spirit, stratospheric intellectual standards, a razor-sharp intuition for the truth, and a boundless imagination. It has been a pleasure and a privilege to work with him, to hear his constant feedback on my own imperfect thoughts. I hope in the future I can honor a tiny portion of his gifts to me by teaching others.

As a PhD student I have been privileged to stand on the shoulders of other both brilliant and kind intellectual giants, without whom this work would never have been. First and foremost, my thesis work owes an unpayable intellectual debt to the work of Graham Cairns-Smith. His unconventional thoughts about the first life on earth were the catalyst for this work on self-replication. I am flattered and grateful for his continued support in the form of visits, talks, and letters during his retirement.

No one was more honest about the rigors of the PhD process and a life in science than Paul Rothmund. As human and as good a friend as Paul has been, he also been someone to aspire to be like. Simply, Paul is a whiz, and a big friendly intellectual giant. I am excited about everything he will do next.

I first visited Gerald Joyce's office about three years ago with a crazy idea and the hope I might convince him to pay attention. Luckily Jerry is a really nice guy. Since then, he's generously served on my committee, travelling from San Diego to do so, and has been a real cheerleader for my work in the field he has been so essential to. I appreciate the time he has taken to look into the details of my work and offer very particular suggestions. Jerry has kept me afloat in the difficult sea of chemistry.

I would never have had any success in many important experiments without Bernard Yurke, a frequent Winfree lab visitor from Lucent. Bernie has answered my stupid questions on everything from solder, which I couldn't even spell before I got here, to unit conversions, to the Navier-Stokes equation, and, as someone also interested in self-replication, has been a constant sounding board and steady voice of reason.

I'd also like to thank the rest of my committee, Jehoshua Bruck, Yaser Abu-Mostafa, and Zhen-Gang Wang. Their kindness and breadth of perspective, questions, and objections have contributed significantly to what I have to say.

There are many people who generously contributed both their hands and minds to the work described in this thesis. I'd particularly like to thank Robert Barish for his collaboration in our work nucleating DNA nanostructures. Rob has spent his undergraduate years working the Winfree lab, and his sheer genius for nanotechnological engineering, his breadth of knowledge, maturity, and the fun of working with him have continually inspired and buoyed me. I'd also especially like to thank another extremely talented undergraduate who did all the hard work in one summer on

another project contained in this thesis, Christina Wright. I wish her the best in all of her future endeavors, but I know she won't need my wishes.

I am thankful to Ho-Lin Chen, Dave Zhang, Sung-Ha Park, and Jonathan Seitel, my collaborators on work during my PhD not included here, for their enthusiasm and talents. They have all been a pleasure to work with.

I thank Bernie, Rizal Hariadi, and Patrick O'Neill for generously spending their time teaching me about fluorescence microscopy. I thank Saurabh Vyawahare of the KNI Microfluidics Foundry for his help and guidance, and for fabricating microfluidic devices used in my experiments.

There are so many other brilliant people I have had the privilege of associating with and who have taken the time to give me feedback and point me in the right direction. I thank Andrew Turberfield and Andrew Ellington for several wonderful conversations about evolution, Ashish Goel and Deborah Fygenon as collaborators and close friends in the field of DNA nanotechnology, and Zhen-Gang Wang and Richard Flagan for pointing me back, several times, in the direction of chemistry convention. Donald Cohen generously donated his brilliance trying to help me solve impossible problems. Eric Davidson hosted me in his lab for three months; his unwavering vision and impeccable sense of logic about the perfectly illogical world of cell biology have been an inspiration. Niles Pierce and his group have worked closely with ours. His lab's friendship has been a boon, and I can only aspire to be a tenth as effortlessly focused and clear as he is. I'd like to thank Frances Arnold for her faith in me, and for giving me the courage to be ambitious.

The Winfree lab has been a constant source of foment and rigorous intellectual exercise. For their constant advice, companionship, and criticisms – valid all – I'd like to thank my colleagues Georg Seelig, David Soloveichik, Matt Cook, and Peng Yin, as well as Hareem Tariq and Si-Ping Han, longtime guests in the Winfree lab.

I would never be at Caltech without having already had such wonderful scientific advisors to teach and encourage me. Thank you to Gerald J. Sussman, Tom Knight, Boris Katz, and Joan Schwartz. Thank you also to Karolyn Yong, the world's most amazing administrative assistant. Everything is easy in her hands.

Caltech is sometimes a hard place to be a woman, but it would be impossible without Candace Rypisi. I thank her for drying many of my tears, feeding me as a hungry graduate student, and for creating a place of refuge. Candace has moved on to great things, but I know the Caltech Women's Center will dearly miss her. Nadine Dabby has also made my life happier when things were difficult. Her whimsical socks have been a vital source of comic relief. I would never be here without Matthew Malchano, one of the best people I know, who helped me apply here and made me come.

I thank Marc Kamionkowski for his constant support and for his infinite patience and love. Thanks Mom and thanks Dad, for putting up with so much, and for letting your daughter be lost to California.

I dedicate this thesis to my grandmother, Rose Porrino. Her unconditional love is the most important thing to me in the world.

Abstract

How life began is still a mystery. While various theories suggest that life began in deep sea volcanic vents or that a world where life consisted predominantly of RNA molecules preceded us, there is no hard evidence to give shape to the chain of events that led to cellular life.

Perhaps the fundamental enigma of our origins is how life began to self-replicate in such a way that evolution could produce Earth's "endless forms most beautiful." With the exception of biological organisms, we have no examples of self-replicating, evolving chemical systems, despite an extensive research program with the goal of identifying them.

In this thesis, I construct a chemical system that preliminary evidence suggests is capable of the most basic self-replication and evolution. The system uses no enzymes or biological sequences, can support and replicate a combinatorial genome, and is completely autonomous. There are no fundamental obstacles to the replication by this system of much more complex sequences or to open-ended evolution.

The design of the system is inspired by the work of Graham Cairns-Smith, who has proposed that life began with clay. Clays are tiny layered crystals; some clay crystals can contain one of several different patterns of atoms or molecules in each layer. The choice of patterns for the layers could be viewed as a sort of genome: it would be copied as the clay grew, and if the crystal broke, each new piece would inherit its pattern from the old piece and could replicate it in the same manner. If some patterns of layers grew and reproduced faster than other patterns, crystals with faster-growing patterns would be selected for.

Instead of the atoms or small molecules of which clay consists, I use molecules consisting of 4 to 6 interwoven, synthetic DNA strands called DNA tiles to design crystals that in principle can replicate and evolve as Cairns-Smith imagined. While the choice of construction material was influenced by ease of use – in contrast to clay crystals, DNA tile crystals have been previously characterized and are easy to synthesize and image in the laboratory – the choice was fundamentally made because DNA tile monomers are programmable, allowing us to create novel crystal morphologies rationally.

The crystals I construct, termed "zig-zag ribbons," contain a sequence of information ("a genome") in each row. Growth of the ribbon adds rows, one at a time, each of which contain an arrangement of DNA tiles that encode the same information sequence as the previous row. Altering the set of "tiles" used to assemble ribbons allows us to alter the alphabets for and the permitted lengths of sequences that can be copied.

I describe how to design tile sets that can replicate genomes with different alphabets and the kind of sequence evolution that is in theory possible with some simple tile sets. Altering the tile set can not only change the kinds of sequences that may be replicated, it can also make growth and splitting more robust. I show how to make changes to the crystals' design to prevent errors during growth and splitting and to reduce the rate of spontaneous generation of new crystals.

It has been previously shown that DNA tile crystallization can be used to perform universal

computation; I show that in theory crystals that can compute can undergo open-ended evolution as they try to produce more and more complex programs to take advantage of available growth resources. This mechanism is simple enough to potentially observe in the laboratory in the near future.

This work suggests that the concept of a self-replicating chemistry is closely related to the concept of a chemistry that can store information and compute. It is only by clearly understanding how chemical systems can transfer and process information that we can hope to understand how self-replication and evolution can occur, and by implication, understand how life might have begun.

Contents

Acknowledgements	iii
Abstract	v
I Introduction	1
1 Preamble	2
1.1 The Origin of Life as a Field of Study	3
1.2 Self-Replication in Computer Science	5
1.3 Programmable Chemistry	5
1.4 Evolution in Simple Systems	7
1.5 Contents and Contributions	8
2 Self-Replication and Evolution of DNA Crystals	10
2.1 Introduction	10
2.2 Replicating Information with DNA Crystals	11
2.3 Simple Evolution: The Royal Road	13
2.4 Selection of Regular Languages	14
2.5 Acceptable Error Rates for DNA Tile-Based Evolution	16
2.6 Conclusions	17
II High Fidelity Replication	19
3 The Design of Tile Sets That Prevent Spontaneous Generation	20
3.1 Introduction	20
3.2 The Zig-Zag Tile Set	26
3.3 The Self-Assembly Model	27
3.4 Thermodynamics of Zig-Zag Assemblies	30
3.5 An Asymptotic Bound on Spurious Nucleation Rates	35
3.6 Numerical Estimates of Spurious Nucleation Rates	40
3.6.1 Spurious Nucleation Rates at Steady State	42
3.6.2 Stochastic Simulations for Estimating Nucleation Rates Before Steady State is Achieved	49
3.6.3 Nucleation of Long Ribbons	50

3.6.4	Expected Effectiveness in Practice	50
3.7	Conclusions	51
3.7.1	Nucleation of Algorithmic Self-Assembly	51
3.7.2	Detection of a Single DNA Molecule	52
4	Preventing Spontaneous Nucleation in the Laboratory	54
4.1	Introduction	54
4.2	Materials and Methods	55
4.3	Results	56
4.4	Conclusions	59
5	Copying Information Accurately by Using Proofreading	65
5.1	Introduction	65
5.2	Experimental Design	67
5.3	Simulation	68
5.4	Results	69
5.5	Discussion	71
III	Evolution	78
6	The Potential for Evolution of Complex Programs	79
6.1	Introduction	79
6.2	Zig-Zag Ribbon Evolution	82
6.3	Dynamics of Crystal Evolution	84
6.4	A Zig-Zag Ribbon Metabolism	85
6.5	Evolution of Universal Sensing and Response	90
6.6	Discussion	95
7	Evolution of Complex Crystal Sequences from Simple Parts	97
7.1	Introduction	98
7.2	A Growth Model for Zig-Zag Crystals	100
7.3	Finite Cellular Automata	102
7.4	Evolution of Zig-Zag Crystals Encoding Binary Cellular Automata	105
7.5	Evolution of Zig-Zag Crystals Encoding Reversible Binary Cellular Automata	108
7.6	Simulation	109
7.7	Conclusions and Open Questions	109
	Bibliography	113

Part I

Introduction

My dear sir, in this world it is not so easy to settle these plain things. I have ever found your plain things the knottiest of all.

Herman Melville, *Moby Dick*

Chapter 1

Preamble

The origin of life may be the last great mystery in biology. While it is not a new problem, it is one for which we have few plausible ideas, let alone well-supported ones. Yet it is of central importance. If “nothing in biology makes sense except in the light of evolution” [Dob64], then knowing what cells evolved from is fundamental.

The scientific problem of life’s origins came about when Darwin proposed that life evolved through the natural selection of fitter individuals. Darwin provided compelling evidence that life changed over time that this change had a scientific explanation. By implication, evolution had a beginning, which in the same way might be a subject of scientific study.

Unfortunately, the nature of Darwin’s theories were such that they did not tell us *how* science might explain such a beginning. In an 1871 letter to Joseph Hooker, Darwin suggested that life may have begun in a “warm little pond, with all sorts of ammonia and phosphoric salts, lights, heat, electricity, etc. present, [so] that a protein compound was chemically formed ready to undergo still more complex changes.” [Dar88]. Such speculations about life’s origins were neither testable nor refutable.

One-hundred and twenty-five years later our ideas about the origin of life are still for the most part neither testable nor refutable. While we know more about some basic questions than Darwin did—What was the Earth like when life emerged? When did life emerge?—we are still in many ways stuck on them [CJAG95, MAM⁺96, Org98a]. How life arose from inanimate materials remains a mystery [Org98b, Org98a]. It’s not even clear what kind of evidence we might hope to find to verify or refute our hypotheses.

Nevertheless, some progress has been made. Since Darwin, many other scientists have produced many ideas about what might have happened. We can also largely eliminate many possibilities for our origins [Eig71, Smi70]. In the past 15 years or so an experimental discipline of attempting to systematically study candidates for an origin of life has arisen (e.g. [Joy96, SBL01, GC04, FC06]), which is giving rise to a field where one can rigorously study and discuss how life *could* have arisen. In the same way that fundamental studies of physical systems under the extreme conditions that might have existed in the early universe inform our understanding of the origins of the cosmos, it seems reasonable to believe that investigating the potential of chemical systems to replicate, evolve, and speciate could inform a rigorous discussion of how life as we know it began.

We are still far from a definitive chronology of how life began. We are even still far from a well-supported idea of how life *could* have begun. To agree on how life originated, we must both find compelling evidence that life did originate as we imagine and a clear and compelling way

to eliminate competing hypotheses. While we cannot eliminate any material from a role in the origin of life, we know something about the physical requirements for self-replication. Thus, a rigorous search for our origins must take into account the following: *The ability to correctly copy its genome is essential to any system capable of self-replication and evolution.* The question of how self-replication can come about was studied first as a logical problem, in computer science, and the results were clear: the ability to compute makes the task of self-replication relatively easy [vNAWB66, Sip97]. We might even speculate, therefore, that the ability of a physical system to compute is closely related to its ability to self-replicate.

So what kind of chemistry can do computation? The notions that the modern cell is essentially an information processing organ [Bra95] and the desire to build computing elements the size of molecules have stimulated a study of how molecular information processing is best accomplished. While much of this work is relatively new, both fundamental theoretical results and a diversity of excellent experimental results have been achieved. Studies of the thermodynamics of computation [BL85] and of the capabilities of abstract chemical systems [Mag97] inform the limits of this chemistry. Work on self-assembly, in which information is processed to guide the assembly of a supramolecular structure [RPW04] and on the assembly of molecular circuits [BGBD⁺04] have suggested important possibilities for the kinds of chemical systems that can process information.

A chemistry that is a candidate for the origin of life must not only be able to self-replicate but also to evolve a genome over time that is more complex than the one with which it began. We know even less about how such a process might happen than we do about how self-replication might arise. While we are now starting to know a little about fitness landscapes of RNA [FS98, CODS04, GGS05] and protein [HS96, CK06] structures and catalysts, and about how changes in developmental programs can cause changes at the organism level [Dav06, Car06], an understanding of how large-scale transitions in biological evolution come about is not understood at the molecular level [SS95]. Further, work on evolution of computer programs has suggested that most systems that can self-replicate and evolve are not capable of open-ended evolution [BMP⁺00]. Thus, understanding what is necessary for a self-replicating system to evolve into a more complex life form is an important part of the question of the origin of life, and still a research problem.

This thesis adopts the perspective that the origin of life is a problem of information processing. It is both a theoretical exploration and a physical demonstration of one way that a system satisfying the important attributes for life can be constructed from molecules. I hope this work is a demonstration of the power of this unconventional approach. By using a programmable chemistry for our explorations, we can investigate the information processing capabilities of a whole class of chemistries (Although much, of course, is lost in translation.) It is my hope that such broad-based chemical explorations, conceptually guided by computer science, will bring fresh insight to the uniquely challenging question of how life began.

1.1 The Origin of Life as a Field of Study

To ask how life started, it is necessary to decide what life is and what it is not. While a definition for life is in many ways impossible to pin down, some reasonable requirements for life are that it be autonomous, self-replicating, and subject to Darwinian evolution [BRC04]. These requirements are not only clear, but at the current time well directed: Self-replication and evolution are fundamental to any life, and yet fascinating because they get at the “fuzzy” region between life and non-life that we don’t understand.

What kinds of chemistry can autonomously self-replicate and evolve? Except for biology, no one has shown conclusively that any chemical system can do so. Nevertheless, many scientists have had ideas about what kinds of chemistry *might* do so, and therefore what the first life could have been like. One of the first to put forth an idea (after Darwin) was Aleksandr Oparin, a Russian scientist who suggested that cell-like coacervates could have formed spontaneously, then grown and divided much as modern cells do [Opa03]. As a graduate student of Harold Urey, Stanley Miller showed that a mixture of highly reducing compounds then thought to be the components of the early earth would form organic molecules when jolted with electricity [Mil53]. The “primordial soup” theory suggests that these organic molecules would be concentrated, and there would gradually form complex and self-replicating entities, much as Darwin imagined. While the idea that life emerged as soft matter in a liquid environment is widely accepted, Graham Cairns-Smith has suggested a very different scenario, that clay crystals could propagate particular defects or arrangements of layers during growth. If these clay crystals were ripped apart by forces such as erosion, the copied defects or layers would be propagated by each of the pieces [CS66].

James Watson and Francis Crick’s discovery of the structure of DNA [WC53] suggested the possibility that a molecule alone might be able to store and replicate information¹. Soon after this work, L. S. Penrose used DNA as the inspiration for an article about the necessary features of a self-replicating system [Pen58]. This article followed a demonstration of many of these principles: his construction with his son Roger Penrose of a self-replicating system of wooden blocks [PP57].

Others have sought to reconstruct with chemistry such a single-molecule self-replicating system. Frances Crick and Leslie Orgel proposed that RNA might have been the molecule whose replication preceded current cells [Cri68, Org68]. The discovery that RNA had catalytic activity bolstered the idea that an RNA world, where RNA both stored genetic information and performed the catalytic activity necessary for replication, was a step on the path to the modern organism. Günter von Kiedrowski showed that enzyme-free replication of a short template was possible with a short palindromic RNA sequence that could template the assembly of two halves of itself, provided those halves were properly chemically activated [vK86, SvK98].

The ability to generate random sequences of RNA and select for their function [Joy89, RJ90] led to a new search for an RNA “replicase,” a particular RNA molecule that could catalyze its own replication. An exhaustive search of RNA sequence space has uncovered some amazing catalysts that can fully or partially replicate themselves given suitable parts. Gerald Joyce’s group recently demonstrated that a particular palindromic RNA sequence could catalyze its own synthesis by ligating together two smaller oligonucleotides [PJ02]. David Bartel, in work stemming from his earlier work with Jack Szostak [BS93], has produced a ribozyme that can copy a template of up to 14 base pairs [JUL⁺01]. In all these cases, molecules replicate themselves (or portions of themselves), satisfying part of the requirement for a life-like system that could have been our origins. However, it is not yet clear whether or how far any of these systems could evolve.

Could other non-template based molecules or assemblies have replicated themselves? Reza Ghadiri’s group demonstrated that a peptide can catalyze its own assembly from two smaller peptides [LGMKSa96]. The self-replication of membrane vesicles has also been demonstrated [WWF⁺94]. Unfortunately, the case for non-trivial evolution of these systems is less encouraging than the evolution of RNA sequences.

Thus, the question of “what kinds of chemistry are capable of Darwinian evolution?” is es-

¹Interestingly, Schrödinger made the suggestion that an aperiodic crystal stored our genetic information in his 1944 book [Sch44].

pecially interesting because it turns out to be hard. Evolving chemistry is all around us in the form of modern biology: cells, viruses, prions, and the like, but thus far it hasn't been possible to demonstrate a self-replicating chemistry without using biological enzymes, which are themselves the product of biological evolution.

1.2 Self-Replication in Computer Science

While making a chemical system that can self-replicate is a research problem, writing a self-replicating computer program is easy. A program that copies itself can generally be written extremely succinctly. In English, such a program might be²

```
Write this sentence twice, the second time in quotes. ‘‘Write this sentence twice,
the second time in quotes.’’
```

This sentence has an analogue in any Turing universal programming language. Further, the recursion theorem [Sip97] tells us that writing a program capable of replication *and* any other computable function is also feasible in any such language.

To what extent is such an observation applicable to the physical world? Among John von Neumann's last scientific contributions, published in a volume completed by Arthur W. Burks [vNAWB66], was the invention the cellular automaton, a model of local computation that takes place on a physical grid. Von Neumann invented this model of computation to study self-replication. Inside the rules of this cellular automaton, he created a self-replicating machine that could both perform universal computation and reproduce itself. von Neumann's machine was complex: it contained 29 different parts (states), which produced a lookup table for each rule consisting of 29⁵ entries, and the process of replication takes so much space that to my knowledge no computer has yet simulated it [Pes95, BH00]. But while the machine was complex, it was also designed in a model that attempted at some level to recapture the physical world.

Von Neumann's machine was complex in part because it could not only self-replicate, but could also construct arbitrary patterns of states according to instructions that were later replicated. But self-replicating cellular automata that can allow limited evolution can be much simpler. Very small 2-dimensional automata [RACP93] and a tiny program in a model of computation without fixed geometry called a graph automaton [TKM02] have been discovered. Since universal computation can be implemented with extremely simple cellular automata [Coo04], it seems reasonable to expect that a small cellular automaton capable of both universal construction and replication exists.

1.3 Programmable Chemistry

The ease with which self-replication can be implemented in a computer program suggests that if there was a chemical system capable of even basic information processing, it might be able to execute a one of the simple self-replicating, evolving computer programs computer scientists have discovered. Such a system would itself be able to self-replicate and evolve.

What is a programmable chemistry? Biology seems to be the clearest example—when one inserts a piece of DNA with a particular sequence (a program) into a bacterium, the bacterium alters its behavior in accordance with the content of the added DNA. Informally, we'll say that a

²This sentence is not my creation but I cannot reliably locate its origins.

programmable chemistry is one that can simulate an abstract model of computation, for example, digital logic circuits or a Turing machine.

What can chemistry compute? An investigation of the limits of our ability to build computing machines inspired some of the first investigations into programmable chemistry. Charles Bennett suggested that DNA transcription, translation, and replication could be viewed as computation, and that such computation was several orders of magnitude more efficient than silicon circuit computation, almost optimally efficient (according to the rules of thermodynamics) [BL85]. Later, Leonard Adleman experimentally demonstrated that computation by a very different method, the manipulation of a library of DNA using standard molecular biology techniques, could be used to efficiently generate solutions to combinatorial search problems [Adl94]. Paul Rothemund showed chemical processes could perform not just complex, but universal computation: a combination of DNA strands and restriction enzymes could, in principle, simulate a Turing machine [Rot96]. Computation in chemistry without enzymatic reactions or polymers is also possible—it can simulate both arbitrary boolean circuits [Mag97] and stochastic chemical kinetics in performing universal computation [CSWB07]. Thus, the simplest kinds of chemistry can, in theory, perform any computation.

In practice, many chemical systems appear to be capable of universal computation. Erik Winfree suggested that universal computation could be embedded in the process of self-assembly [Win96], and he and others implemented a set of molecules which, with the correct programming, could perform these computations [WLWS98]. Progress has also been made constructing systems that do not use polymerization or crystallization to compute and therefore contain only a finite number of species. Recent work by Seelig et al. showed that DNA, without enzymes, could implement simple boolean circuit functions, and that other than the limits of DNA sequence binding specificity, this functionality could arbitrarily be expanded []. Media as diverse as reaction-diffusion systems [SKS96], (pseudo-)rotaxanes [CBLS97], quantum-dot cellular automata [AOT⁺99], and synthetic peptides [AG04] have been shown to also perform simple computations in the laboratory, suggesting that the phenomenon of computation may be widespread in chemistry.

We copy information in crystals using Winfree’s method of algorithmic self-assembly [Win96, WLWS98, RPW04], algorithmic self-assembly. Algorithmic self-assembly has its roots in the tiling problem, the question of whether a given set of shapes can tile the plane [Wan61]. By using a set of “Wang tiles,” squares with colors on each side that can be laid adjacent only to squares with matching sides, Hao Wang showed that a version of this problem in which some tiles are already added to the plane was undecidable [Wan62]. Later, Robert Berger showed that even if no tiles were added to the plane to start, answering the general question of whether a set of tiles could tile the plane was also undecidable [Ber66]. Winfree’s work demonstrated the computational universality of self-assembly by showing that the chemically plausible assembly of molecular analogues of Wang tiles was also computationally universal. Under conditions where only cooperative attachments of blocks (by either one strong or two weak crystal bonds) was allowed, the assembly of a 2-D lattice could in principle simulate a 1-D blocked cellular automaton, and therefore could also perform universal computation. In his construction, a row of the 2-D crystal contains the state of the automaton at one time step, with growth of the crystal in the forward direction advancing the computation. Growth is nucleated from a structure which provides the initial state and prevents backwards growth. When two edges of the block must match the existing tiling to attach to it—these two edges can be considered the “input” states, and the remaining two edges on the block the “output” of that computation. Since growth can continue indefinitely, arbitrarily long computations

can be performed.

Algorithmic self-assembly can be implemented in practice using DNA double crossover molecules (synthetic assemblages of short oligo-nucleotides) [FS93, LYK⁺00, MSS99, YPF⁺03, CSK⁺04, HCL⁺05] as Wang tiles. The DAO-E DNA double crossover molecules [FS93] used in the experiments in this thesis consist of several short oligonucleotides (20 to 80 base pairs). When annealed, the oligonucleotides assemble into a brick-like structure due to a preference for Watson-Crick complementarity. The “core” of a DNA double-crossover molecule is double stranded, and therefore not reactive with other DNA molecules, but base pairs on each of four ends of two helices in the molecule are single stranded. The four single stranded edges are analogous to the colors of the Wang tiles—tiles generally bind only to tiles or crystals with complementary sticky end sequences. A set of many such molecules can therefore be viewed analogously to a set of Wang tiles. Work by Mao et al. [MLRS00a], Rothmund et al. [RPW04], Barish et al. [BRW05], and others have demonstrated that the mechanism Winfree described can indeed guide the crystallization of DNA double crossover molecules and produce aperiodic patterns in crystals that are the result of computations.

In this thesis, I describe how this system for creating “molecular jigsaw puzzles” can be used to construct 1-D crystals (crystals that are several tiles wide, but cannot change their width once nucleated). Crystals elongate via the ordered addition of DNA tiles, one row (of fixed width) at a time. The set of molecules is designed so that a crystal can grow only by adding tiles that copy the existing arrangement of tiles. Thus, each layer of the crystal contains the same piece of information. Sustained growth produces many copies of the sequence.

A self-replicating system requires not only accurate computation to copy a sequence, but also the ability to separate the sequence and its copy or copies. In practice, separation of two copies is riddled with problems, including non-separation of the original genome and its copy (end-product inhibition) and destruction of the copied genome. One solution to this problem is to make many copies of the genome before separation is required, so that even if the separation process is error-prone, at least a few copies of the genome remain intact. In this thesis, rough mechanical forces is used to split the crystals into pieces, each of which can then grow new copies of the sequence. While some pieces will be damaged, enough should remain intact to replicate the original genome.

1.4 Evolution in Simple Systems

Thus, this work describes how to autonomously replicate a sequence of information that is embedded within a crystal. We will also show how in some environments, crystals bearing different sequences can compete with each other. It is therefore possible to ask whether crystal evolution could lead to the creation of complex crystal forms, and as such be a candidate for an ancestor of biological organisms.

Many biochemists believe that the answer to this question for almost any system capable of Darwinian evolution is “yes”: once Darwinian evolution is possible in an autonomous system, more and more complex life would be forthcoming (e.g. [CS06]), and therefore the creation of a self-replicating chemical might be the rate-limiting step in producing life.

Computer scientists have been studying self-replication computer programs for a long time [Sip98]. Yet despite numerous attempts to write programs that can evolve into more and more complex programs, many programs seem to evolve to a certain point and stop [Bed07]. In a few cases, simulations have shown continued adaptation [CWO⁺04], but it’s not clear if that adaptation would continue indefinitely. Further, the level of adaptation and increased complexity are not on the scale

of that seen in biological evolution. It is still unclear what the properties of a system capable of open-ended evolution would be [BMP⁺00]. In the absence of examples of evolving chemical systems, one might be tempted to assume that evolution in a self-replicating chemical system might also be difficult.

Yet in many respects, biology satisfies the intuitive requirements for what one might call “open-ended evolution”: from some system simple enough to arise spontaneously, life today has become more complex than any machine we can currently build [EG00, VAM⁺01, DS86, Int07]. How did this happen? One might choose to make some observations about life that might bias it toward the occurrence of at least some evolution. First, biological organisms can replicate a combinatorial variety (any DNA sequence imaginable) of genotypes. Second, a genotype produces a clear phenotype—a set of associated molecules that enable metabolism, protection, further replication, etc. Third, phenotype has a clear effect on the ability to replicate—an ineffective enzyme can prevent digestion of a vital resource, or a sloppy method of copying DNA might produce dead offspring. Lastly, complicated phenotypes can allow an improvement in fitness in complex environments—the production of a new enzyme could allow the digestion of a new sugar or an increased ability to outwit competitors for a scarce resource. An ability to find adaptations that beget selective advantage seems to be essential for open-ended Darwinian evolution.

1.5 Contents and Contributions

The chapters that follow represent an assemblage of published work or work in preparation to be published.

The remainder of the introductory part of this thesis, Chapter 2, explains in detail how in principle a DNA tile system for self-replication and evolution can be constructed. It shows DAO-E DNA tiles [FS93] that can be assembled to form a crystal that, as it grows, copies a sequence, and how the “life-cycle” that Cairns-Smith envisioned for clay crystals can be adapted to DNA tile crystals. The second part of this work describes what it means for such a crystal to evolve, and describes some example chemistries (sets of DNA tiles) and environments in which simple evolution could occur.

The second part of this thesis describes how we overcame major technical challenges, both theoretical and experimental, to achieving self-replication and evolution with DNA crystals. All of these challenges relate to the fact that for evolution to reach an optimum on a particular fitness landscape, information replication has to occur accurately [Eig71, EMS88]. In the context of DNA crystals this means that we need to ensure three things: that information within crystals are copied correctly, that nucleation of crystals containing random sequences occurs rarely, and that crystals don’t lose information when broken.

Techniques to reduce errors during growth of 2-D DNA crystals by redesigning the molecular components of the crystal were known previously [WB04, RSY05, CG05]. Chapter 3 addresses the second requirement described in the previous paragraph, by describing how a similar type of redesign can in theory drastically reduce the rate of nucleation of crystals while only marginally reducing their growth rate. Chapter 4 describes laboratory experiments that verify that this redesign technique reduces crystal nucleation rates. Chapter 5 describes how we adapted techniques to prevent growth errors in 2-D crystals to improve the fidelity of information copying in our crystals. The experiments described in that chapter show that this adaptation does work and confirms further predictions about growth “proofreading” that were too difficult to perform in two dimen-

sions: that increasing the amount of redundancy introduced in the redesign process decreases error rates.

Part 3 describes my work on crystal evolution. The first two chapters ask the question “What is all this stuff good for?” since it’s not immediately obvious to the biochemist schooled in catalysis, that DNA crystals actually *do* anything that might beget them selective advantage. In Chapter 6, I describe one thing they can do: perform universal computation, and how that implies that DNA crystals are capable of highly non-trivial evolution. In an environment where resources for growth can change, well-adapted crystals, like biological life forms, can sense which resources they’re likely to have access to and use those for growth. In Chapter 7, I show that a similar mechanism of selection for complex crystals in some environments can take place even with a very small set of tiles. These results indicate that we can begin doing experiments to look for open-ended evolution of DNA crystals in the near future.

Contributions

A version of Chapter 2 has been published as:

Rebecca Schulman and Erik Winfree. Self-Replication and Evolution of DNA Crystals. In *Advances in Artificial Life, 8th European Conference*, pages 734–743. Springer-Verlag, 2005.

Erik and I wrote this paper. Many of the conceptual ideas in this paper are due to Erik.

A version of Chapter 3 has been submitted for publication. I am first author on this work, and Erik Winfree is my coauthor. I proved all the theorems in the paper and did the simulations and the analysis. Erik and I together decided on the content of and wrote the paper.

A version of Chapter 4 has been submitted for publication. I am first author on this work, and Erik Winfree is my coauthor. I performed all the experiments described in this paper and Erik and I both designed the experiments and did the analysis.

A version of Chapter 5 is in preparation for submission. This work will be authored by Christina Wright, myself, and Erik Winfree. Christina performed the experiments and designed and did the programming required for the simulations. I designed the experiments and all the molecules used and wrote key parts of the simulation engine. I wrote the paper with help from Erik.

A version of Chapter 6 has been accepted for publication as

Rebecca Schulman and Erik Winfree. How Crystals that Sense and Respond to Their Environments Could Evolve. *Natural Computing*, 2007.

The ideas are mine, refined by discussion with Erik. I wrote the paper with Erik.

I conceived of the ideas in Chapter 7, and refined them in conversations with Erik Winfree. I wrote the chapter.

Chapter 2

Self-Replication and Evolution of DNA Crystals

Abstract

It is widely accepted that Darwinian evolution is responsible for the complexity and adaptability seen in modern biology. However, the mechanisms by which evolving organisms adapt to their environment are not well understood. A central roadblock to studying evolution is the dearth of physical systems in which Darwinian evolution can be studied; a tractable synthetic system for replication and evolution would facilitate the study of how physical selection pressures lead to evolutionary change. A chemical self-replicator might also be used to evolve solutions to problems in chemistry or nanotechnology. If such a system were simple enough, it could also shed light on how self-replication emerged spontaneously at the origin of life.

2.1 Introduction

In 1966, Graham Cairns-Smith proposed a simple mechanism by which polytypic clay crystals could replicate information in the absence of biological enzymes [CS66, CS88]. Polytypic clay crystals contain discrete layers, each of which contain molecules of a particular identity or orientation. A cross-section of the crystal therefore contains an information-bearing sequence. Cairns-Smith proposed that crystal growth extends the layers, copying the sequence (the crystal's genotype). Occasionally, physical forces break a crystal apart. Because crystals replicate their genotype many times during growth, splitting of a crystal can yield multiple pieces, each containing at least one copy of the information-bearing sequence. Cycles of growth and fragmentation might therefore cause each sequence to be exponentially amplified.

We propose a method of self-replication that works by a similar process of growth and fragmentation. Instead of replicating sequences in clay, we suggest that this process could also be used to amplify sequences in algorithmic DNA crystals. DNA crystals are composed of DNA tile monomers [FS93]. Different types of DNA tiles can be designed to assemble via programmable rules [Win96]; a typical DNA crystal is assembled from several tile types. As in Graham-Smith's conception, DNA crystals can contain a sequence that is copied during growth, in this case a linear arrangement of DNA tile types (Figure 2.1a). Unlike most types of clay crystal growth, DNA

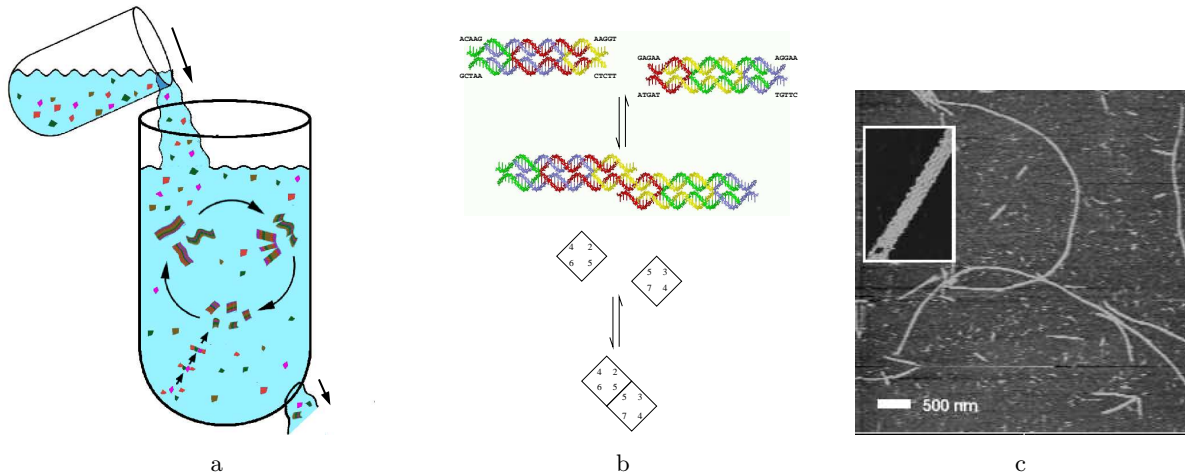


Figure 2.1: **DNA crystals.** (a) The DNA crystal life cycle. The materials required for growth are constantly replenished. Crystals die when they are flushed out of solution in an exit stream. (b) Tiles with complementary single-stranded sticky ends can attach by hybridization. For convenience, DNA tiles may be represented as square tiles; tiles with the same side labels correspond to molecules with matching sticky ends. (c) Atomic force microscopy image of DNA crystals formed by the molecules shown in Figure 2.2b. At higher resolutions (inset), individual tiles can be discriminated.

crystal growth is tractable in the laboratory and occurs at time scales (hours) that are suitable for experimental investigation.

In Section 2.2, we describe in more detail how crystal evolution works and introduce the components of DNA crystals and a model of the growth process. The examples in Sections 2.3 and 2.4 illustrate how DNA crystals can copy arbitrary amounts of information and how in particular environments, this information affects the replication rate. In Section 2.5, we describe techniques for increasing the accuracy of replication.

2.2 Replicating Information with DNA Crystals

The DNA crystals we propose consist of DNA tile monomers [FS93] which can attach to other tiles in a programmable fashion: each of the four sides of the DNA tile has a short single stranded portion which can hybridize with the complementary strand of another tile (Figure 2.1b). DNA tiles can assemble into 2-dimensional crystals [WLWS98] and can be programmed to form other structures, such as thin ribbons (Figure 2.1c). A wide variety of DNA tile crystals have been synthesized [MSS99, LYK⁺00, CBG⁺04, HCL⁺05].

Under slightly supersaturated conditions [Win98], the attachment of a DNA tiles to a growing crystal is only energetically favorable if it occurs cooperatively, i.e., by the formation of two or more sticky end bonds. The attachment of a tile to a crystal performs a step of a computation in the sense that a unique tile (among many possible in solution) may attach at a particular growth location. With an appropriate choice of tiles, DNA tile assembly can perform universal computation [Win96].

The zig-zag crystal shown in Figure 2.2a is formed from the tiles shown in Figure 2.2b. Matching rules determine which tile fits where. When a zig-zag crystal is added to a solution of free tiles

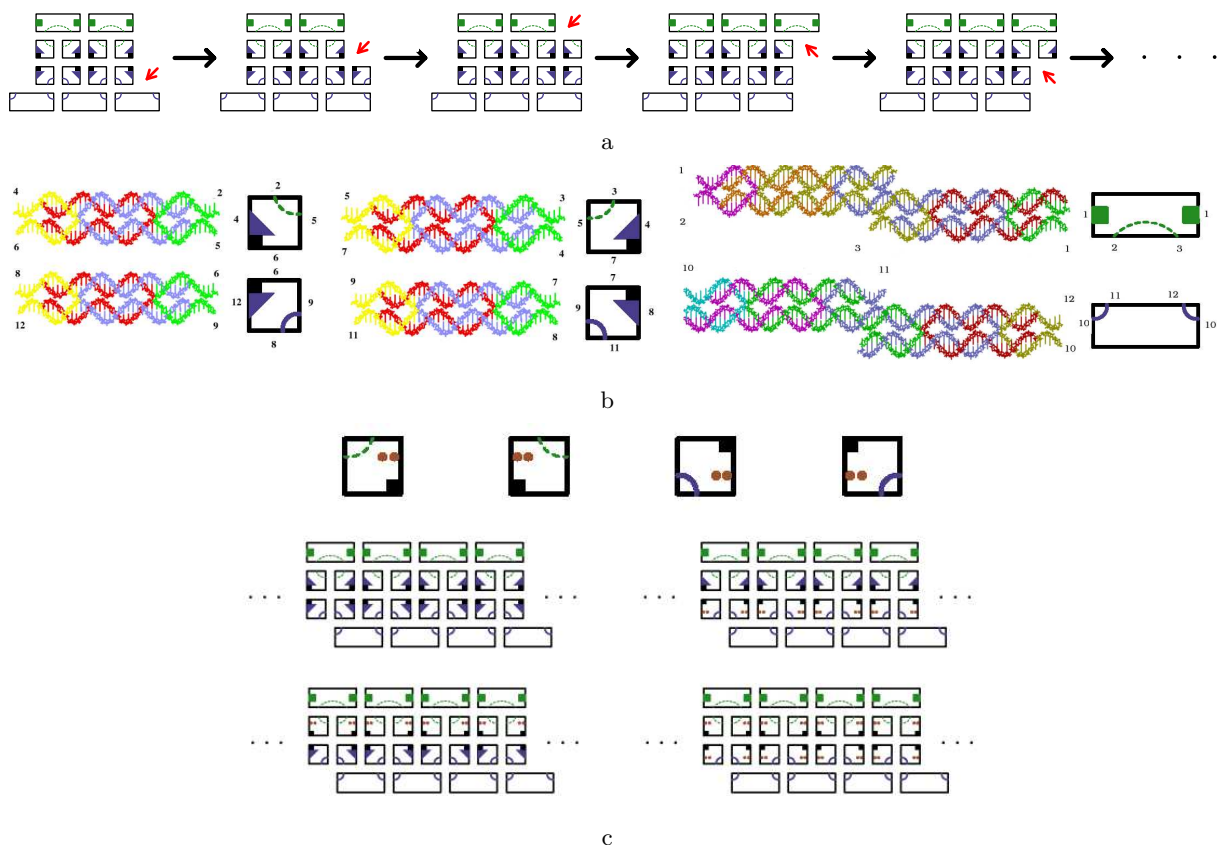


Figure 2.2: **The zig-zag tile set.** (a) A zig-zag assembly. Two alternating tile types in each row enforce the placement of the double tiles on the top and bottom, ensuring that under slightly supersaturated conditions, growth occurs in a zig-zag pattern. Although only growth on the right end of the molecule is shown here, growth occurs simultaneously on both ends of the molecule. At each step, a new tile may be added at the location designated by the small arrow. (b) The basic zig-zag tile set consists of six molecules (tile types). Each square and rectangle shown is a logical representation of the molecule shown to its left. By convention, tiles cannot be rotated. The tiles shown here have unique bonds that determine where they fit in the assembly: each label has exactly one match on another tile type. The logical representations of DNA tiles have the same connectivity as DNA tile molecules but are rotated 45° and have a different aspect ratio. (c) The tile set shown in Figure 2.2b forms only one type of assembly. A tile set consisting of the tiles in (b) and the four tiles shown here allows four types of assemblies to be formed. The vertical column of each type contains a different 2-bit binary sequence.

under slightly supersaturated conditions, growth is constrained to occur in a zig-zag pattern by the requirement that each tile addition must form two or more sticky end bonds, as shown in Figure 2.2a. It is easy to confirm that under such conditions, there is always a unique tile that may be added on each end of the ribbon.

Zig-zag crystals are designed so that under slightly supersaturated conditions, growth produces one new row at a time, and continued growth repeatedly copies a sequence. The requirement that a tile must attach by two bonds means that it must match both its vertical neighbor (another tile that

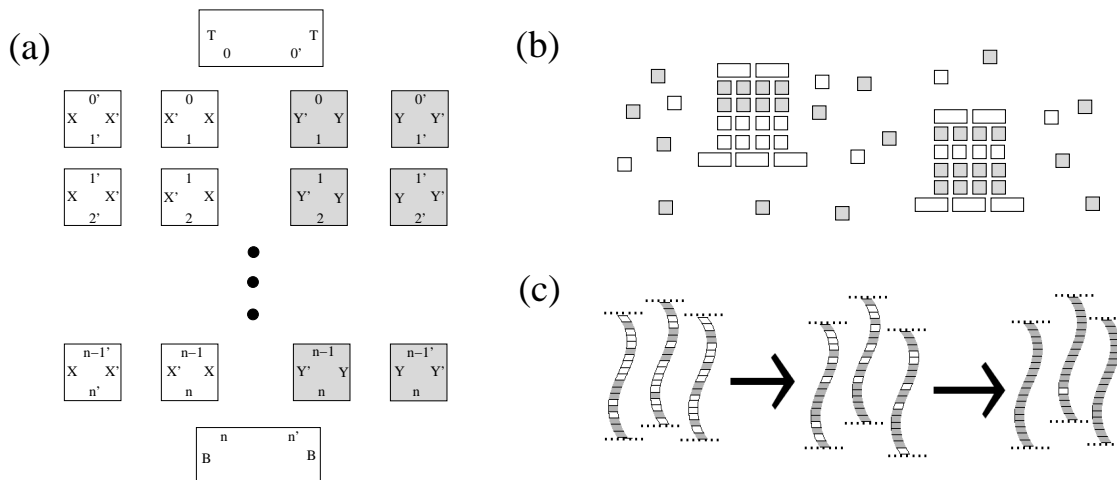


Figure 2.3: **The royal road tile set.** (a) The royal road tile set consists of four tiles for each of n sequence positions, two for propagating an X bit and two for propagating a Y bit. Two boundary tiles are also used. 2^n different sequences can be copied with this tile set. (b) When more Y than X tiles are present, sequences containing more Y tiles tend to grow faster. (c) As growth progresses, sequences containing mostly Y tiles become more and more common. Each sequence shown represents an assembly consisting of many copies of the illustrated sequence.

is part of the new column being assembled), and its horizontal neighbor (in a previously assembled row). Several tiles might match the label on the vertical neighbor, but because tiles must make two correct bonds in order to join the assembly, only a tile that also matches the label on the horizontal neighbor can be added. Therefore, the tile being added in the new column must correspond to the one in the previous column. As a result, information is inherited through templated growth. The set of tiles formed by adding the tiles in Figure 2.2c to those shown in Figure 2.2b can propagate one of four strings. Additional tiles may be added to the set of tiles in Figures 2.2b and 2.2c to create a tile set that copies one of 2^n sequences of width n . We will later discuss tile sets in which an unbounded amount of information can be copied.

The growth of a zig-zag DNA crystal increases the number of copies of the original information present in the ribbon, but does not change the rate at which new copies of the sequence are produced. The rate of copying can be sped up by tension elongation force which causes crystals to break. With each new crystal that is created by breakage, two new sites become available to copy information. Repeated cycles of growth and breakage exponentially amplify an initial piece of information. Occasionally, a tile matching only one bond rather than two will join the assembly, resulting in occasional copying errors, which are also inherited. If errors happen during copying, which they will under almost any achievable condition [Win98], and crystals with particular sequences grow faster than others, then evolution can occur.

2.3 Simple Evolution: The Royal Road

An artificial selection experiment involves both an environment (a set of resources for growth, their chemistry and the ambient physical conditions) and an initial population of organisms. Here the environment includes a set of DNA tiles; we assume all reactions take place under physical

conditions where copying errors occur very occasionally. The set of DNA tiles determines the set of sequences which may be copied and the “chemistry” of the system, i.e., the rules which tiles bind to each other¹. A particular arrangement of DNA tiles is the information that is propagated in these experiments, the genotype; it is the organism being evolved. The phenotype of a sequence is its replication rate in the given environment. In this section we describe a tile set that allows many kinds of sequences to grow; a selection pressure results from physical conditions in which tile concentrations differ for each tile type.

A DNA crystal can grow only when it comes in contact with matching tiles. We assume assembly occurs in a well-mixed reaction vessel where the higher the concentrations of tiles of the type that may be legally added the vessel contains, the more quickly such contact occurs. Therefore, a simple selection pressure results from a difference in concentration between tile types used to copy the sequence information: assemblies with sequences containing tiles present at high concentrations will grow and reproduce faster than assemblies with sequences containing tiles present at very low concentrations.

A tile set in which one of two bits can be propagated at each of n sequence positions is shown in Figure 2.3a. Let X_i and Y_i be the two tile types that can be propagated at position i . If Y_i is present in solution at a concentration higher than that of X_i , as suggested by Figure 2.3b, the fitness landscape for this selection resembles the simplest case of a well-studied problem in genetic algorithms, the “royal road” [MFH92]. Here, the growth rate increases monotonically with the number of Y_i s in the sequence s . So long as the Y_i tiles remain more common in solution, sequences containing only Y_i tiles will quickly dominate (Figure 2.3c).

2.4 Selection of Regular Languages

Section 2.3 illustrated how tile concentration can create a selection pressure, causing some sequences to grow faster than others. While this is a simple selection pressure to understand, the adaptation that occurs is also simple. In this section we describe how a single tile set allows for the replication of an infinite number of sequences and how sequence constraints imposed by the tile set can provide more interesting selection pressures.

In the previous example, the “chemistry” of the tile set determined the length of sequences that could be copied, and which tiles could be used in which position of the sequence. Here we consider evolution when the tile set “chemistry” allows only certain sequences to be copied, but these sequences may be arbitrarily long. In particular, the tile set environment allows copying only of sequences that are accepted by a particular finite state machine.

A finite state machine is an abstract device that can perform a computation requiring only a fixed amount of memory. It consists of a set of states and rules describing how to transition between states as each character of input is received. Computation begins in a prescribed state. When the inputs have all been received, the current state is either in an accept state, in which case the the input is accepted, or a reject state. Figure 2.4a shows a simple finite state machine (along with the tiles that implement the transition steps of the machine) which detects whether the number of

¹Our choice of terminology reflects the observation that whether a self-replicator is made from clay, biological polymer, or other material, the chemistry of the specific elements involved determines the evolutionary landscape. As an example, the chemistry of nucleic acids can make some sequences hard to copy. Certain sequences fold up or bulge [Joy87], making copying of those sections more difficult. Here, the constraints are not on how a sequence folds, but on how its elements fit together: the tile set similarly determines the evolutionary landscape.

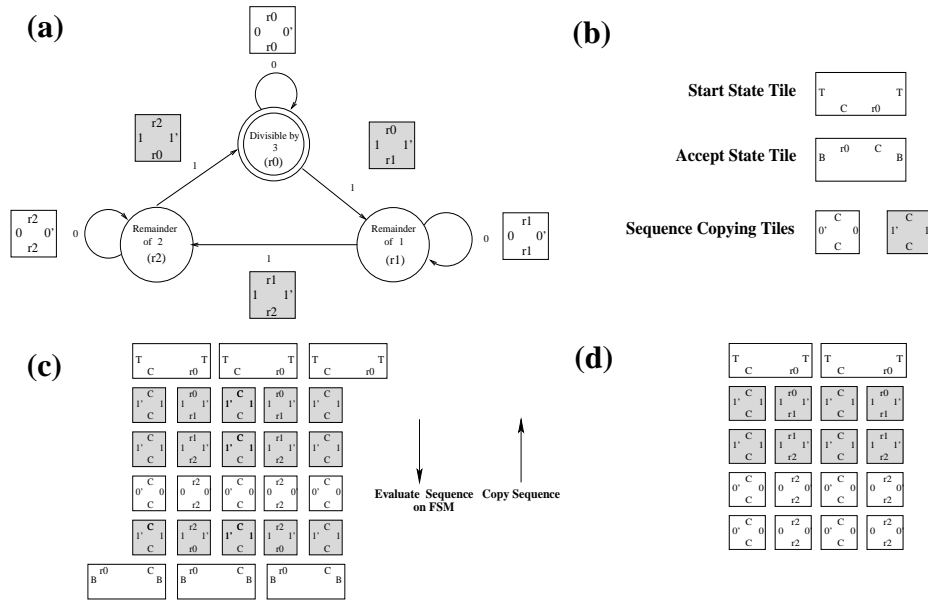


Figure 2.4: **Selection of sequences with particular numbers of logical 1s.** (a) A diagram of a finite state machine that can determine whether a binary sequence contains a number of 1s that is divisible by 3. The double circled state is both the start and the accept state. (In general these states are not the same.) The tiles shown can be used with the tiles in (b) to follow the instructions of the machine during tile assembly. (b) Additional tiles needed to complete the tile set in (a). The construction shown in (a) and (b) can be generalized to any finite state machine. (c) An assembly encoding a sequence accepted by the machine in (a). Evaluation ends in an accept state, so a bottom tile may be added and assembly can continue. (d) An assembly encoding a sequence not accepted by the finite state machine in (a). Because execution of the finite state machine ended with a state other than the accept state, assembly cannot continue.

ones in a binary sequence input is divisible by three.

The self-assembly of DNA sequence [Adl94] and tile [Rei97] alphabets can generate the set of sequences accepted by a given finite state machine, also known as a regular language. Accepted sequences can be of any length. In contrast to the tile sets described in Section 2.3, where the top and bottom sides of a tile encode the position in the fixed-length sequence where the tile can be added, the top and bottom sides of the tiles in Figure 2.4a encode the state of the machine as it processes each character of the sequence being copied.

Constructing a tile set that copies only inputs accepted by a given finite state machine is straightforward. Each possible transition between states is encoded as a single tile (Figure 2.4a). The left and right sides of the tile encode the input, the top side encodes the state that machine is in before the input is received and the bottom side encodes the state that the machine transitions to after the input has been received. The top boundary tile encodes the start state and a bottom boundary tile encodes each accept state (Figure 2.4b). Another set of tiles copies a sequence that has been accepted by the machine. These tiles have only one state on their bottom and top sides, and encode the same sequence bit on their left and right sides.

During growth down the crystal², assembly evaluates the sequence according to the finite state machine’s rules. If the machine ends in an accept state, a bottom tile can bind to the site and upward growth can begin (Figure 2.4c). If the machine is not in an accept state, no bottom tile exists which matches the growth front, and growth stops (Figure 2.4d). Thus, only sequences which are accepted by the machine will continue to be replicated. These sequences will be the ones that are selected for.

More complex selection pressure results if the crystals grown in this tile set environment are moved to an environment containing tiles that accept a different language of sequences. For example, crystals grown using the tiles shown here might be moved to a mixture containing tiles that allowed only sequences with a number of ones that is divisible by 5 to grow. Only sequences with a number of ones divisible by 15 could survive in both environments.

2.5 Acceptable Error Rates for DNA Tile-Based Evolution

Several experimental studies have shown that DNA tiles can process information through cooperative binding [MLRS00b, RPW04], which suggests that DNA tiles should be able to copy information in the manner that we describe. However, these studies show that errors very often occur during algorithmic assembly [RPW04]. This is a concern because a low error rate is vital to the design of a self-replicator. If the error rate exceeds an error threshold [Eig71], genetic meltdown occurs and sequences become totally random. In this section we describe how in principle, it should be possible to decrease the error rate below any relevant error threshold.

Errors during assembly occur when a tile binds to a growing assembly by fewer than two bonds, an event called an unfavorable attachment. A mismatch error, an unfavorable attachment that only partially matches the adjacent tiles, causes an error in replication (Figure 2.5a). Additionally, in the absence of a pre-existing crystal, a series of unfavorable attachments occasionally produces a full-width crystal with a random sequence, an event called spontaneous nucleation.

Both these kinds of errors can be analyzed using a reversible model of DNA tile self-assembly based on the physics and chemistry of DNA hybridization [Win98]. Prior work on the robustness of algorithmic self-assembly in this model can be adapted in order to show that, at a moderate cost of tile set complexity and assembly speed, mismatch error rates can be made as small as is desired. “Proofreading” tile sets implement the same logic of an original tile set but assemble more robustly, dramatically reducing mismatch error rates without significant slow-down [WB04, CG05, RSY05]. The general idea of proofreading is to redundantly encode each element of sequence. When the proofreading method is applied to the zig-zag tile set (Figure 2.5b), correct tile additions are stabilized by additional tiles in the same block that encode the same sequence element, whereas several incorrect additions instead of just one are needed to propagate a sequence element incorrectly (Figure 2.5c). Error rates decrease exponentially as larger blocks of proofreading tiles are used [WB04].

Similar error correction techniques also exist for the prevention of spontaneous nucleation errors. Like other crystallization processes, the rate at which spontaneous nucleation of growing zig-zag

²Growth on the left side of the zig-zag crystal in Figure 2.4c reads the sequence elements backward, and evaluates the finite state machine in reverse. While running the finite state machine shown in Figure 2.4a backward accepts the same set of states as running the machine forward, for other machines there may be non-determinism when the machine is run in reverse. A step may be possible that cannot lead to the start state, leaving an uncompleted assembly. Assemblies corresponding to tile sets of this type will grow mostly in the direction where the finite state machine is evaluated in the correct direction. Alternatively, it is also possible to replace this tile set with an equivalent tile set that can grow only in the forward direction [Win06], at the cost of tile set complexity.

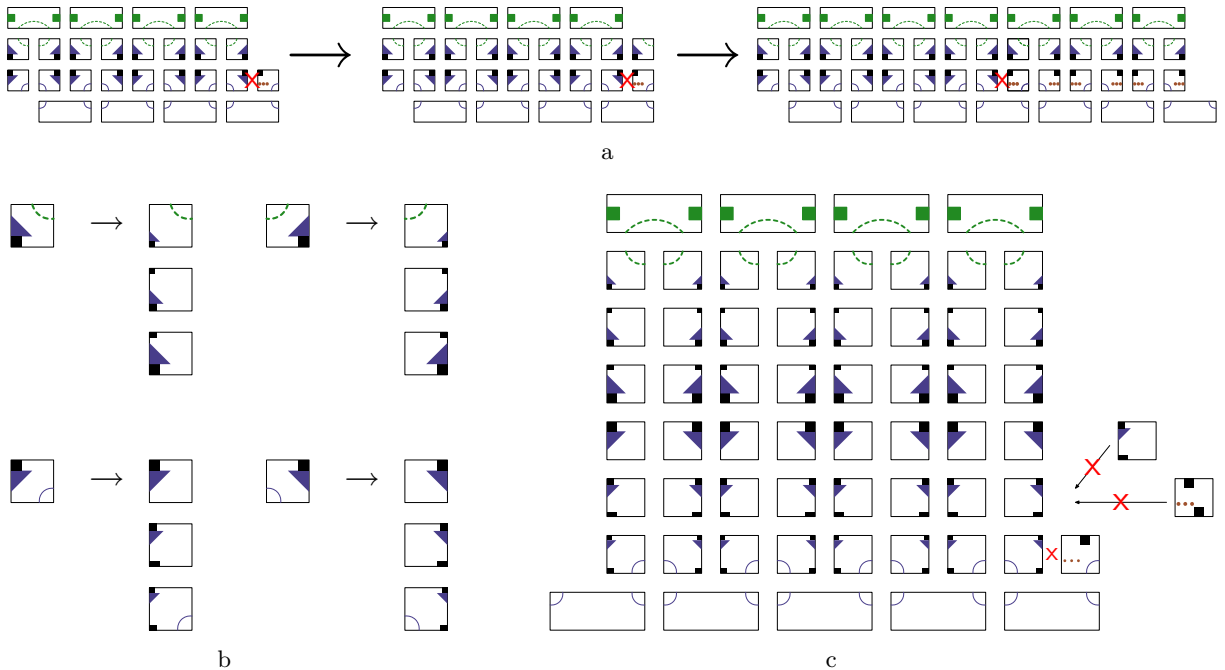


Figure 2.5: **Proofreading for zig-zag assembly.** (a) Kinetic trapping is the major cause of mismatch errors in DNA tile assembly. When a tile attaches to an assembly by only one side, it forms a low energy bond and usually dissociates quickly. However, if a second tile attaches to the assembly adjacent to the first tile before the first tile can dissociate, the first tile may be trapped. The sequence is therefore copied incorrectly. Further growth will propagate the incorrect sequence in subsequent columns. (b) A “proofreading” transformations of the four zig-zag middle tiles in Figure 2.2b to a set of tiles that can in principle copy the same information more robustly. (c) Zig-zag assembly of the original sequences using the transformed tile set. When an incorrect tile attaches to the assembly, either the tile must fall off and be replaced by the correct tile, or further errors are necessary in order to continue growth.

assemblies occurs is dependent on the energy of the critical nucleus, the smallest assembly of tiles that tends to grow into a large crystal rather than melt. For zig-zag crystals, this critical nucleus is a small assembly that contains both a top and bottom boundary tile. By increasing the minimum width of an assembly that can contain both these tiles, it is possible to increase the energy of the critical nucleus. For example, the rate of spontaneous nucleation of the zig-zag tiles in Figure 2.3a decreases exponentially with the width of the crystal in tiles [SW05a]. We expect that the same qualitative result applies to the more complex tile sets described in this paper.

2.6 Conclusions

To study the physical principles of Darwinian evolution, we propose a physical system based on DNA crystals in which a combinatorial variety of genotypes can be faithfully replicated and a genotype can direct a behavior or other measurable parameter that can be subject to selection. DNA crystals are simple, containing no biological parts, and can be programmed to replicate an

infinite variety of genotypes. The ability to program the interactions between tiles allows us to induce selection pressures which favor the growth of assemblies with interesting properties. Error correction techniques exist which can lower the replication error rate as much as is required to avoid genetic meltdown, at the cost of a small amount of additional complexity.

Part II

High Fidelity Replication

O Nature, and O soul of man! how far beyond all utterance are your linked analogies! not the smallest atom stirs or lives on matter, but has its cunning duplicate in mind.

Herman Melville, *Moby Dick*

Chapter 3

The Design of Tile Sets That Prevent Spontaneous Generation

Abstract

Algorithmic self-assembly, a generalization of crystal growth processes, has been proposed as a mechanism for autonomous DNA computation and for bottom-up fabrication of complex nanostructures. A ‘program’ for growing a desired structure consists of a set of molecular ‘tiles’ designed to have specific binding interactions. A key challenge to making algorithmic self-assembly practical is designing tile set programs that make assembly robust to errors that occur during initiation and growth. One method for the controlled initiation of assembly often seen in biology is the use of a seed or catalyst molecule which reduces an otherwise large kinetic barrier to nucleation. Here we show how to program algorithmic self-assembly similarly, such that seeded assembly proceeds quickly, but there is an arbitrarily large kinetic barrier to unseeded growth. We demonstrate this technique by introducing a family of tile sets for which we rigorously prove that, under the right physical conditions, increasing the size of the tile set by a constant amount exponentially reduces the rate of spurious nucleation. Simulations of these ‘zig-zag’ tile sets suggest that under plausible experimental conditions, it is possible to grow seeded crystals in just a few hours such that less than 1 percent of crystals are spuriously nucleated. Simulation results also suggest that zig-zag tile sets could be used for detection of single DNA strands. Along with prior work on constructing tile sets that are robust to assembly errors during growth, this work is a step toward understanding how algorithmic self-assembly can be performed with low error rates without a significant reduction in assembly speed.

3.1 Introduction

Molecular self-assembly is an emerging low-cost alternative to lithography for the creation of materials and devices with sub-nanometer precision [WMS91, Leh93]. Whereas top-down methods such as photolithography impose order externally (e.g., a mask with a blueprint of the desired structure) bottom-up fabrication by self-assembly requires that this information be embedded within the chemical processes themselves.

Biology demonstrates that self-assembly can be used to create complex objects in this way. Organisms produce sophisticated and functional organization from the nanometer scale to the

meter scale and beyond. Structures such as virus capsids, bacterial flagella, actin networks, and microtubules can assemble from their purified components, even without external direction from enzymes or metabolism. This suggests that spontaneous molecular self-assembly can be engineered to create an interesting class of complex supramolecular structures. A central challenge is how to create a large structure without having to design a large number of unique molecular components.

Algorithmic self-assembly has been proposed as a general method for engineering such structures [Win96] by making use of local binding affinities to direct the placement of molecules during growth. The binding of a particular molecule at a particular site is viewed as a computational or information transfer step. By designing only a modest number of molecular species, which constitute the instructions or program for how to grow an object, complex objects can be constructed in principle [RW00, CRW04, SW04]. The implementation of algorithmic self-assembly requires controlled nucleation: An assembly that grows from the right nucleus executes the instructions for assembly in the correct order, while uncontrolled nucleation leads to a spectrum of undesired products. The primary concern of this paper is how to engineer molecules that ensure self-assembly begins with controlled nucleation. We address this question theoretically, using a model that is commonly used to study crystallization [LK97], but which incorporates the particularities of algorithmic self-assembly.

To motivate the model we use, we first describe a specific molecular system that can implement algorithmic self-assembly experimentally. DNA double crossover molecules [FS93] and related complexes [LYK⁺00, MSS99, YPF⁺03, HCL⁺05, CSK⁺04] (henceforth, “DNA tiles”) have the necessary regular structure and programmable affinity to implement algorithmic self-assembly. Simple periodic [WLWS98, LYK⁺00] and algorithmic [MLRS00a, RPW04, BRW05] self-assembly reactions have been realized experimentally. As an example, consider one of the DNA double crossover molecules shown in Figure 3.1, which self-assembles from 4 strands of synthetic DNA. The DNA sequences are designed such that the desired pseudoknotted configuration maximizes the Watson-Crick complementarity [See90, DLWP04]. Since the energy landscape for folding is dominated by logical complementarity more so than by specific sequence details, it is possible to design similar double crossover molecules with completely dissimilar sequences. To date, nearly 100 different molecules of this type have been synthesized.

Interactions between DNA tiles are dictated by the base sequences of each of four single-stranded overhangs, termed ‘sticky ends,’ which can be chosen as desired for each tile type. Tiles assemble through the hybridization of complementary sticky ends. The free energy of association for two tiles in a particular orientation is assumed to be dominated by the energy of hybridization between their adjacent sticky ends. The hybridization energy is favorable when complementary sticky ends bind, but negligible or unfavorable for non-complementary sticky ends. The DNA tiles shown crystallize into sheets via the binding of sticky ends to four adjacent molecules, forming a lattice (Figure 3.1). When multiple tile types are present in solution, each site on the growth front of the crystal preferentially will select from solution a tile that makes the most favorable bonds. Under appropriate physical conditions, a tile that can attach by two sticky ends will be secured in place, while tiles that attach by only a single sticky end usually will be rejected due to a fast dissociation reaction. We call these “favorable” and “unfavorable” attachments, respectively.

The design of an algorithmic self-assembly reaction begins with the creation of a tile program and its evaluation in an idealized model of tile interaction, the abstract tile assembly model (aTAM) [Win98]. A DNA tile is represented as a square tile with labels on each side representing the four sticky ends. Polyomino tiles with labels on each unit-length of the perimeter can

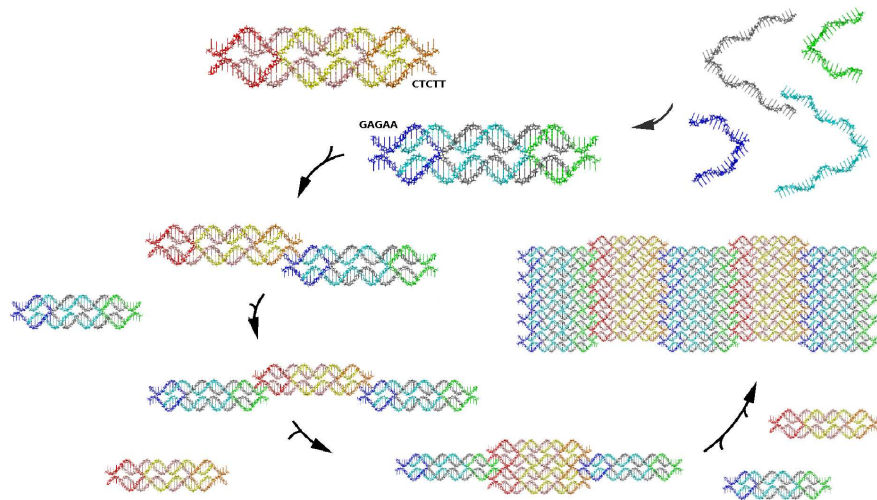


Figure 3.1: A DNA double crossover molecule and its assembly into a 2D crystal

be used in addition to square tiles, since it is possible to generate the corresponding DNA structures. A tile program consists of a set of such tiles, the strength with which each possible pair of labels binds, a designated seed tile, and a strength threshold τ . Under the aTAM, growth starts with a designated assembly of tiles (usually just the seed tile) and proceeds by allowing favorable attachments of tiles to occur. That is, tiles may be added where the total strength of the connections between the tile and the assembly is greater than or equal to the threshold τ . At a given step, any allowed attachment may be performed. Addition of tiles is irreversible. An example of a structure that can be constructed using algorithmic self-assembly, a Sierpinski triangle, is shown in Figure 3.2a. Beginning with the seed tile, assembly in the aTAM will result in the growth of a V-shaped boundary that is subsequently (and simultaneously) filled in by “rule tiles” that obtain their input from their bottom sides and present their output on their top sides. The four rule tiles for this self-assembly program consist of the four cases in the look-up table for XOR. The assembly of these tiles therefore executes the standard iterative procedure for building Pascal’s triangle mod 2. Tile sets for the construction of a variety of desired products have been described [Win96, LL00, RW00, ACGH01, CRW04, AGKS04], including a tile set capable of universal construction [SW04].

In contrast to assembly in the aTAM, the assembly of DNA tiles is neither errorless nor irreversible. Further, the assembly of DNA tiles may not start from a seed tile. In recent experimental demonstrations of algorithmic self-assembly [RPW04, BRW05], between 1% and 10% of tiles mismatched their neighbors and only a small fraction of the observed crystals were properly nucleated from seed molecules. Following [SLPW04], Figure 3.2b illustrates how unseeded nucleation and unfavorable attachments can lead to undesired assemblies.

To theoretically study the rates at which errors occur, we need a model that includes energetically unfavorable events. The kinetic Tile Assembly Model (kTAM) [Win98] describes the dynamics of assembly according to an inclusive set of reversible chemical reactions: a tile can attach to an assembly anywhere that it makes even a weak bond, and any tile can dissociate from the assembly at a rate dependent on the total strength with which it adheres to the assembly (Figure 3.2c). The kinetic tile assembly model is a lattice-based model, in which free tiles are assumed to be well mixed, and effects within the crystal such as bending or pressure differences are ignored. The kTAM has been used to study the trade-off between crystal growth rate and the frequency of mismatches (errors) in seeded assemblies [Win98]. In principle, the rate of some errors can be reduced by assembling crystals more slowly. Analysis of assembly within the kTAM suggests that it is also possible to control assembly errors by reprogramming an existing tile set so as to introduce redundancy. ‘Proofreading tile sets’ [WB04, CG05, RSY05, SW05c] transform a tile set by replacing each individual tile with a $k \times k$ block of tiles, exponentially reducing seeded growth errors with respect to the size of the block.

In this paper we propose a method of transforming a tile set to control nucleation errors. We prove that in principle this method exponentially reduces the rate at which assemblies without a seed tile grow large (unseeded growth), while maintaining the rate of growth that starts from a seed tile and proceeds roughly according to aTAM (seeded growth). To do so, a tile set must satisfy two conflicting constraints: when assembly begins from a seed tile, it must proceed quickly, whereas assembly that starts from a non-seed tile must overcome a barrier to nucleation in order to continue.

How is it possible to have a barrier to nucleation only when no seed is present? In a mechanism for the control of 1-dimensional polymerization, found both in biology [SM01, CDVW04]

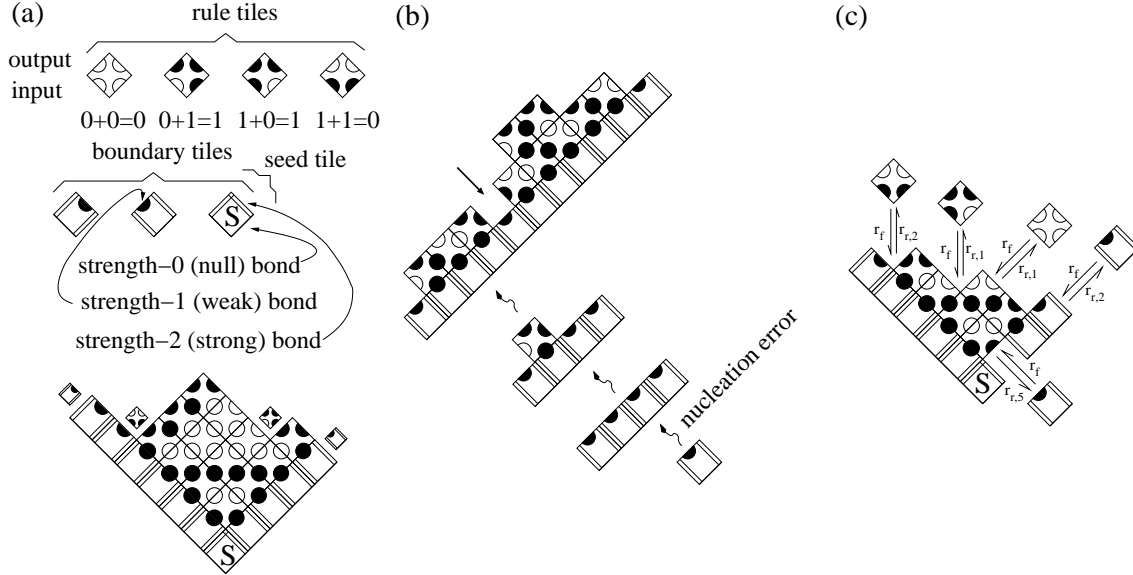


Figure 3.2: **The Sierpinski tile set.** (a) Because DNA tiles are generally not rotationally symmetric, formal tiles cannot be rotated. The lower diagram shows the seeded growth of the Sierpinski tiles according to the aTAM at $\tau = 2$. The small tiles indicate the (only) four sites where growth can occur. When growth begins from a seed, no more than one tile can attach at each location, so assembly results in a unique pattern. (b) Errors resulting from improper nucleation (assembly that does not begin from the seed tile). Tile sets like the one in Figure 3.2a, where two individual tiles can attach to each other with a strength $\geq \tau$ are particularly prone to nucleation errors. Improper nucleation can produce a long facet where an insufficient attachment allows a surrounding block of tiles to attach favorably. Different blocks of tiles may be incompatible, leading to an inevitable mismatch at their intersection. The straight arrow indicates a site where such a mismatch will occur. (c) The rates of tile assembly and disassembly in the kinetic Tile Assembly Model (kTAM). For the growth of an isolated crystal under unchanging tile concentrations, the forward (association) rate in the kTAM is $r_f = k_f[tile] = k_f e^{-G_{mc}}$, while the reverse (dissociation) rate is $r_{r,b} = k_f e^{-bG_{se}}$ for a tile that makes bonds with total strength b . Parameters G_{mc} and G_{se} govern monomer tile concentration and sticky-end bond strength, respectively. A representative selection of possible events is shown here. Attachments with reverse rates $r_{r,1}$ are unfavorable for $G_{mc} > G_{se}$. The kTAM approximates the aTAM with threshold τ when $G_{mc} = \tau G_{se} - \epsilon$. The same set of reactions are favorable or unfavorable in the two models.

and engineering [DP04], a seed induces a conformational or chemical change to monomers, without which monomers cannot polymerize. For example, in spontaneous actin polymerization, it is proposed that a trimer occasionally bends to form an incipient helix that allows for further nucleation [SM01]. The Arp 2/3 protein complex seeds actin polymerization by imitating the shape of an unfavorable intermediate to actin filament nucleation [KAP95]. In two- and three-dimensional systems, condensation of a gas [McD62], crystallization [Mar03], or in general in the Ising model [SJB99], classical nucleation theory [Zet69, DG00] predicts that a barrier to nucleation exists because the free energy of monomer clusters have an unfavorable energy term proportional to the surface area of the cluster (possibly due to interfacial tension or pressure differences with respect to the surrounding solution), and a favorable energy term proportional to the volume of the cluster. Because volume grows more quickly than surface area as clusters grow larger, a supersaturated regime exists where small clusters tend to melt, but above a critical size, cluster growth rather than melting is favored. Aspects of both these methods are combined in some crystalline ribbons or tubes, where growth, initially in two dimensions, is disfavored because of unfavorable surface area/volume interactions, up to the point that the full width ribbon or tube has been formed. A seed structure allows immediate growth by providing a stable analogue to a full-width assembly. Protein microtubules [MBS⁺95] and DNA tubes [MHM⁺04, RENP⁺04, LPRY04] exhibit this type of nucleation barrier.

In this paper we describe a tile set, the zig-zag tile set, that uses this method for the control of nucleation during algorithmic self-assembly. Zig-zag tiles assemble into a potentially long ribbon of predefined width. While a seed tile allows zig-zag ribbons to grow immediately, only full-width boundaries can grow by favorable attachments, so without the seed tile there is a critical size barrier that prevents spurious nucleation. Because it is simple to reengineer the monomers used in self-assembly, by redesigning the tile set it is possible to increase the width and therefore the critical size.

In addition to its intrinsic interest as an engineering scheme for controlling nucleation, the zig-zag tiles solve the aforementioned problem in controlling nucleation during algorithmic growth. With an appropriate seed, zig-zag ribbons can play the same role as the V-shaped boundary in Figure 3.2a. Since rule tiles are not likely to spuriously nucleate on their own under optimal assembly conditions [Win98], once this boundary has set up the correct initial information, algorithmic self-assembly will proceed with few spurious side products.

In Section 3.2, we describe the zig-zag tile set family in detail. In Section 3.3, we explain the mass-action model of self-assembly kinetics that we use to analyze the assembly of zig-zag tiles. In Section 3.4 we analyze thermodynamic constraints on ribbon growth. In Section 3.5, we prove our main theorem, that the rate of spurious nucleation decreases exponentially with the width of the zig-zag tile set. In contrast, the speed of seeded assembly decreases only linearly with width. Thus, for a given volume, we can construct a tile set such that no spurious nucleation is expected to occur during assembly. This illustrates how the logical redesign of molecules can be qualitatively more effective in preventing undesired nucleation than just controlling physical quantities such as temperature and monomer concentration. In Section 3.6, we use both mass-action and stochastic simulations to provide numerical estimates of nucleation rates. These estimates suggest that reasonably sized zig-zag tile sets can be expected to be effective in the laboratory.

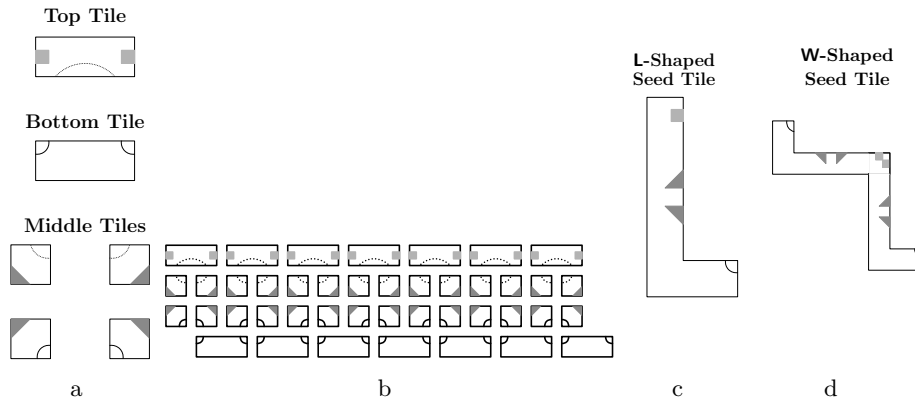


Figure 3.3: **The zig-zag tile set.** (a) The width 4 zig-zag tile set and seed tiles. Each shape represents a single tile. Tiles have matching bonds of strength 1 when the shapes on their edges match. (b) The ribbon structure formed by the width 4 zig-zag tile set (c) The L-shaped seed nucleates linear assemblies. (d) The W-shaped seed tile, with appropriate tiles for vertical zig-zag growth, could nucleate V-shaped assemblies.

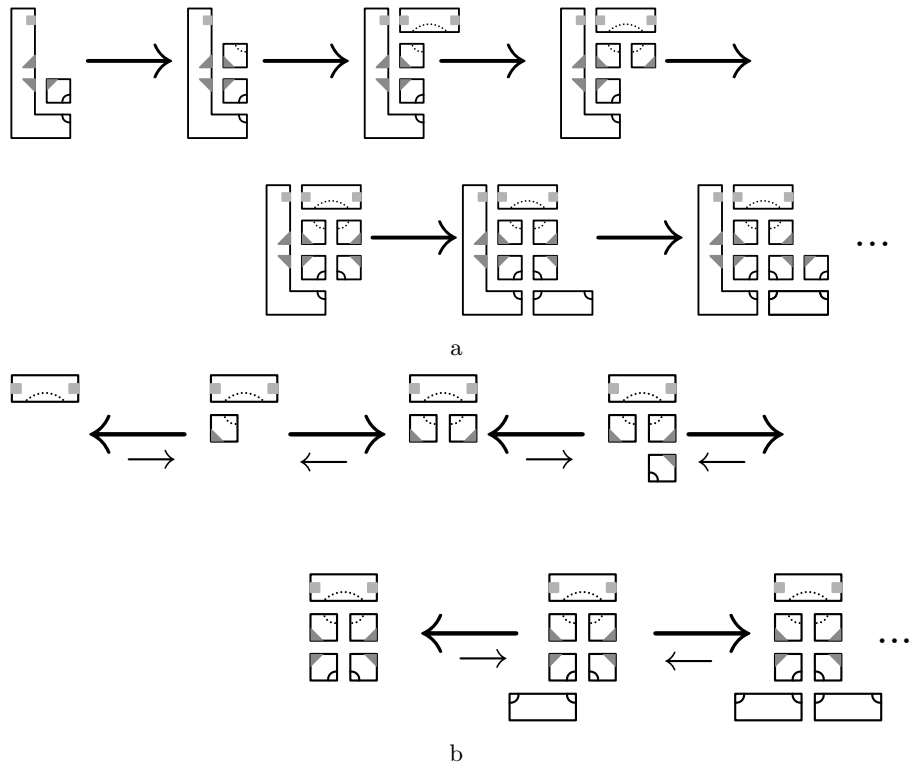


Figure 3.4: **Zig-zag tile set growth.** (a) Seeded growth of a zig-zag tile set in the aTAM. The same growth pattern occurs reversibly in the kTAM when $G_{mc} = 2G_{se} - \epsilon$. (b) Unseeded growth. A possible series of steps by which the tiles could spuriously nucleate in the kTAM

3.2 The Zig-Zag Tile Set

A self-assembly program is a set of tiles that self-assemble into a desired set of tile crystals. A **zig-zag tile set** (Figure 3.3a) of width k assembles to form a periodic ribbon of width k (Figure

3.3b). Zig-zag tile sets of widths $k \geq 2$ can be constructed. A zig-zag tile set includes a **top tile** and a **bottom tile**, each consisting of 2 horizontally connected square tiles. Each of the $k - 2$ rows between the top and bottom tiles contains two unique **middle tiles** that alternate horizontally. The alternation of the two tiles enforces the staggering of the top and bottom tiles. Each tile label has exactly one match on another tile type, so that the tiles cannot assemble to form any other structures held together by sticky end bonds.

The tile set is designed to operate in a physical regime where the attachment of a tile to another tile or assembly by two matching sides is energetically favorable, whereas an attachment by only one bond is energetically unfavorable. In the aTAM, these conditions translate to growth with a threshold of 2. Growth beginning from any tile in the zig-zag tile set goes nowhere in the aTAM—no two tiles can join by at least two bonds. In contrast, growth can proceed from an L-shaped or V-shaped seed tile (Figures 3.3c and 3.3d). Figure 3.4a illustrates the only possible growth path in the aTAM from the L-shaped seed. The staggering of the top and bottom tiles allows growth to continue indefinitely along a zig-zag path. (This growth path is analogous to spiral growth in microtubules but unwrapped onto the plane.) Note that the top and bottom tiles alternately provide the only way to proceed to each successive column. Assemblies that do not span the full width (k tiles) cannot bind both kinds of tiles, and thus cannot grow indefinitely. Growth from a seed tile of less than full width would stall. For example, with a seed tile of width $k - 1$, the top tile could not attach by two bonds to the assembly.

In the kTAM, seeded growth occurs in the same pattern as in the aTAM. Unlike in the aTAM, however, there are also series of reactions that can produce a full width assembly in the absence of a seed tile. The formation of such an assembly is called a spurious nucleation error. An example of such unseeded growth is shown in Figure 3.4b. Under the conditions of interest, some steps in spurious nucleation are energetically favorable, but at least $k - 1$ must be unfavorable before the full width assembly is formed. At this point, further growth is favorable. Spurious nucleation is a transition from assembly *melting*, where assemblies are more likely to fall apart than they are to get larger, to assembly *growth*, where each assembly step is energetically favorable. An assembly where melting and growth are both energetically favorable is called a critical nucleus.

Nucleation theory [Zet69, DG00] predicts that the rate of nucleation is limited by the concentration of the most stable critical nucleus, $[A_c]$. Intuitively, because more unfavorable reactions are required to form critical nuclei in a wider zig-zag tile set, the free energy of A_c and therefore $[A_c]$ at steady state should decrease exponentially with k . This argument is not rigorous, however, because unfortunately the number of critical nuclei also increases with k . The rate of spurious nucleation is proportional to the sum of the concentrations of all these critical nuclei. We will show in the following sections that despite this issue, under many conditions nucleation rates do decrease exponentially with k .

3.3 The Self-Assembly Model

To analyze the process of tile assembly, we formally describe the mass-action kinetic Tile Assembly Model (kTAM). For a given tile set, kTAM describes the set of possible assembly reactions and the dynamics of these reactions. The kTAM has been previously used to analyze the assembly process of other tile programs [RW00, SW04], and is a general framework for understanding algorithmic self-assembly.

A **tile type t** is a unit square, or a polyomino (a finite, connected set of unit squares) with

bond types on each unit side of the tile. The set of possible **bond types** is referred to as Σ . A set of tile types is denoted by \mathbf{T} . A tile (as contrasted with a tile type) is a tuple of a tile type and a coordinate location $(x, y) \in \mathbb{Z}^2$. The set of tiles (all possible tile types in all possible locations) is referred to as T . Tiles cannot be rotated. Tiles that abut vertically or horizontally are **connected** if they have the same labels on the abutting sides. A set of tiles is connected if there is a path of connected tiles between any two tiles in the set.

An assembly A is an equivalence class with respect to translation of a non-overlapping, connected finite set of one or more tiles. The set of assemblies is denoted by \mathcal{A} . A set of tiles \tilde{A} is considered the canonical representation of A if

$$\tilde{A} \langle t, (x, y) \rangle \in \tilde{A} \text{ s.t. } x < 0 \text{ or } y < 0 \text{ and } \exists \langle t, (0, y) \rangle \in \tilde{A} \text{ and } \exists \langle t, (x, 0) \rangle \in \tilde{A}.$$

That is to say, the canonical representation uses a coordinate system such that the assembly just fits in the upper right quadrant of the plane with no negative coordinates. $\tilde{A}(x, y)$ refers to the tile type at coordinates (x, y) in this representation, or 0 if there is no such tile. For an assembly A , $\text{width}(A) = \max_x |y_1 - y_2|$, such that $\langle \mathbf{t}_1, (x, y_1) \rangle, \langle \mathbf{t}_2, (x, y_2) \rangle \in A$. The length of A , $\text{length}(A)$, is defined analogously. The addition relation is defined between an assembly A and a tile t so that $A + t = B$ if and only if \tilde{A} and t are connected but non-overlapping, and $\tilde{A} \cup t$ is a member of equivalence class B . For the attachment of two tiles to each other, we consider the set of tile types \mathbf{T} to be listed in some order, and treat the first tile as an assembly. The addition relation is therefore defined between two tiles t_1 and t_2 only when t_1 comes before t_2 . $t_1 + t_2 = A$ if and only if $t_1 = \langle t, (0, 0) \rangle$, t_1 and t_2 are connected but non-overlapping, and $t_1 \cup t_2$ is a member of equivalence class A . This definition is crafted to correctly count the number of distinct ways in which tiles can attach to each other. For example, two tiles of the same type with the same label on all four sides can attach in exactly four distinct ways, and two tiles of different types that have matching left and right sides respectively, and unique bonds on all the other sides can attach in exactly one way.

If abutting labels on two connected tiles match, these tiles form a **bond**. The standard free energy, G° , of an assembly A is defined as $G^\circ(A) = -bG_{se}$, where b is the number of bonds in the assembly and G_{se} (the sticky end energy) is the unitless free energy of a single bond.

The mass-action kTAM model considers reversible chemical reactions between tiles and assemblies occurring in well-mixed solution. In this paper, we consider all possible accretion reactions: reactions either between two tiles or between a tile and an assembly. In the powered kTAM, single tile concentrations are held constant during assembly (i.e. the reaction is “powered”). In a powered accretion model of self-assembly, a reaction’s rate is dependent on at most one changing concentration, so dynamics are linear. Early in crystallization, the period modeled here, most tiles are still unbound, and similarly most reactions are accretion reactions, so a powered accretion model is a reasonable approximation to a reaction where tile concentrations are not held constant.

Formally, the set of **powered accretion reactions** are

$$R = \{A + t \rightarrow B + t, B \rightarrow A : A, B \in \mathcal{A} - T, t \in T, A + t = B\} \cup \{t_1 + t_2 \rightarrow A + t_1 + t_2, A \rightarrow \emptyset : t_1, t_2 \in T, A \in \mathcal{A} - T, t_1 + t_2 = A\}$$

The appearance of single tiles on both sides of the association reactions and neither side of the dissociation reactions reflects the powered model’s assumptions that the concentration of single tiles remains constant.

Mass-action chemical reaction dynamics [DB02] and the powered accretion reactions define for

each assembly a differential equation that describes the rate of change of the assembly's concentration in time. The concentration of assembly A is denoted by $[A]$. In general, for a chemical reaction $\sum_i n_i S_i \rightarrow \sum_j m_j S_j$ with rate constant k , where $n_i, m_j \in \mathbb{Z}^{\geq 0}$, and S_i are chemical species, mass-action dynamics predict $\frac{d[S_j]}{ds} = k(m_j - n_j) \prod_i [S_i]^{n_i}$, where s represents time. (We use s instead of t because t denotes a tile in this paper.) These dynamics add linearly for multiple reactions.

In the kTAM, each reaction has a forward rate constant k_f that we assume to be the same for all reactions, and a backward rate constant $k_r = k_f e^{-\Delta G^\circ}$, where ΔG° is the difference between the sum of the standard free energies of the reactants and that of the products (where the standard free energy of a single tile is 0). The concentration of all tile types is held at $e^{-G_{mc}}$. (Identical concentrations are considered for convenience only; we show below how our formalism can be extended trivially to treat reactions where species have different concentrations.) Assemblies consisting of more than a single tile have an initial concentration of 0. Thus, for an assembly A at time point s ,

$$\frac{d[A]}{ds} = k_f \left(\sum_{\substack{A+t \rightarrow B+t, \\ B \rightarrow A \in R}} e^{G^\circ(B) - G^\circ(A)} [B] - [A] e^{-G_{mc}} + \sum_{\substack{B+t \rightarrow A+t, \\ A \rightarrow B \in R}} [B] e^{-G_{mc}} - e^{G^\circ(A) - G^\circ(B)} [A] + \sum_{\substack{t_1+t_2 \rightarrow A+t_1+t_2, \\ A \rightarrow \emptyset \in R}} e^{-2G_{mc}} - e^{G^\circ(A)} [A] \right). \quad (3.1)$$

Each term in the first summation is the difference between the rate at which A and a tile react to form a larger assembly B and the rate at which the larger assembly B decomposes into A and a tile. Each term in the second summation is the difference between the rate of formation of A by a reaction where a single tile binds to a smaller assembly B , and the rate decomposition of A into assembly B and a single tile. The terms in the final summation are the rate of formation of A from two single tiles. In the remainder of this paper, we refer to the mass-action kTAM with powered accretion reactions as simply “the kTAM.”

The free energy $G(A)$, in contrast to the standard free energy, reflects both the entropy loss due to crystal formation and the enthalpy gain of assembly. It is defined as $G(A) = G^\circ(A) + \sum_{t \in A} G_{mc}$. Because tile concentrations are constant, the steady-state concentration of an assembly A is given by $[A]_{ss} = e^{-G(A)}$. For an assembly with n tiles and b bonds, this concentration is $[A]_{ss} = e^{(bG_{se} - nG_{mc})}$.

Lemmas 1 and 2 show that this model satisfies detailed balance within $\mathcal{A} - T$. The lemmas also apply to the case, not considered in this paper, where different tile types have different (but constant) concentrations. For a tile t , $S(t)$ is defined as the relative concentration of its corresponding tile type. Unit concentration is $e^{-G_{mc}}$, such that the concentration of tile type \mathbf{t} , $[\mathbf{t}] = S(\mathbf{t}) e^{-G_{mc}}$.

Lemma 1 *For all reaction pairs $A + t \rightarrow B + t$ and $B \rightarrow A$, $k_f [t] [A]_{ss} = k_r [B]_{ss}$, where k_f and k_r are the rates of the respective reactions.*

Proof By definition $k_r = k_f e^{-\Delta G^\circ}$, so that

$$\begin{aligned} k_r [B]_{ss} &= k_f e^{G^\circ(B) - G^\circ(A)} [B]_{ss} \\ &= k_f e^{G^\circ(B) - G^\circ(A)} e^{-G(B)} \\ &= k_f e^{G^\circ(B) - G^\circ(A)} e^{-(G^\circ(B) + \sum_{t' \in B} G_{mc} - \ln(S(t')))} \\ &= k_f e^{-G^\circ(A)} e^{-\sum_{t' \in B} (G_{mc} - \ln(S(t')))}. \end{aligned}$$

Because $A + t = B$,

$$\begin{aligned}
&= k_f e^{-G^\circ(A)} e^{-\sum_{t' \in A} (G_{mc} - \ln(S(t')))} e^{-G_{mc} + \ln(S(t))} \\
&= k_f e^{-G(A)} e^{-G_{mc} + \ln(S(t))} \\
&= k_f [A]_{ss} [t].
\end{aligned}$$

Lemma 2 For reaction pairs $t_1 + t_2 \rightarrow A$ and $A \rightarrow \emptyset$, $k_f[t_1][t_2] = k_r[A]_{ss}$.

Proof

$$\begin{aligned}
k_r[A]_{ss} &= k_f e^{G^\circ(A)} [A]_{ss} \\
&= k_f e^{G^\circ(A)} e^{-G(A)} \\
&= k_f e^{G^\circ(A)} e^{-(G^\circ(A) + 2G_{mc} - \ln(S(t_1)) - \ln(S(t_2)))} \\
&= k_f e^{-G_{mc} + \ln(S(t_1))} e^{-G_{mc} + \ln(S(t_2))} \\
&= k_f [t_1][t_2].
\end{aligned}$$

While only equal strength sticky ends are considered in this work, this proof shows that detailed balance applies to a model of self-assembly with arbitrary sticky end strengths.

3.4 Thermodynamics of Zig-Zag Assemblies

To prove that nucleation rates of zig-zag ribbons decrease exponentially as their widths increase, we would first like to identify the critical nuclei for spurious nucleation. Thermodynamic constraints provide a powerful tool: Because assemblies with unfavorable energies have low steady-state concentrations, we can conclude that such assemblies occur rarely without having to consider rates. (In contrast, assemblies with favorable energies may or may not form quickly, depending upon details of the kinetics; such analyses form the bulk of Sections 3.5 and 3.6.)

We therefore consider the free energy landscape, where each point in the landscape corresponds to a particular type of assembly. Optimal control over nucleation is achieved in a regime where zig-zag growth is favorable, but the growth of less than full-width (thin) assemblies is unfavorable.

Within the kTAM, the energy landscape for assemblies is formally described by the free energy $G(A) = bG_{se} - nG_{mc}$, which can be evaluated directly for any given assembly A . G_{se} and G_{mc} describe the physical conditions for assembly. Recalling that the steady-state concentration of an assembly A is $e^{-G(A)}$, changing G_{se} and G_{mc} can bring the system into two qualitatively different phases. In the melted phase, $G(A)$ is bounded below by $-G_{mc}$ for all A , meaning that no assembly has an appreciable concentration at steady state. In contrast, in the crystalline phase, $G(A)$ can continue to decrease without bound as certain polymeric assemblies become longer and longer—that is, adding a repeat unit to the assembly strictly decreases its free energy¹. Within the crystalline phase, there are regimes where different types of long polymers are favorable or unfavorable, depending on the physical parameters. To ensure that thin polymers do not tend

¹In powered models, since $[A]_{ss} = e^{-G(A)}$, free energies that decrease without bound for longer polymers imply that formal steady-state concentrations correspondingly increase without bound. This may seem odd, but it is not problematic; it reflects that providing unbounded materials can lead to an unbounded accumulation of product, and that longer polymers do not reach steady state within the time during which the powered model is an appropriate model.

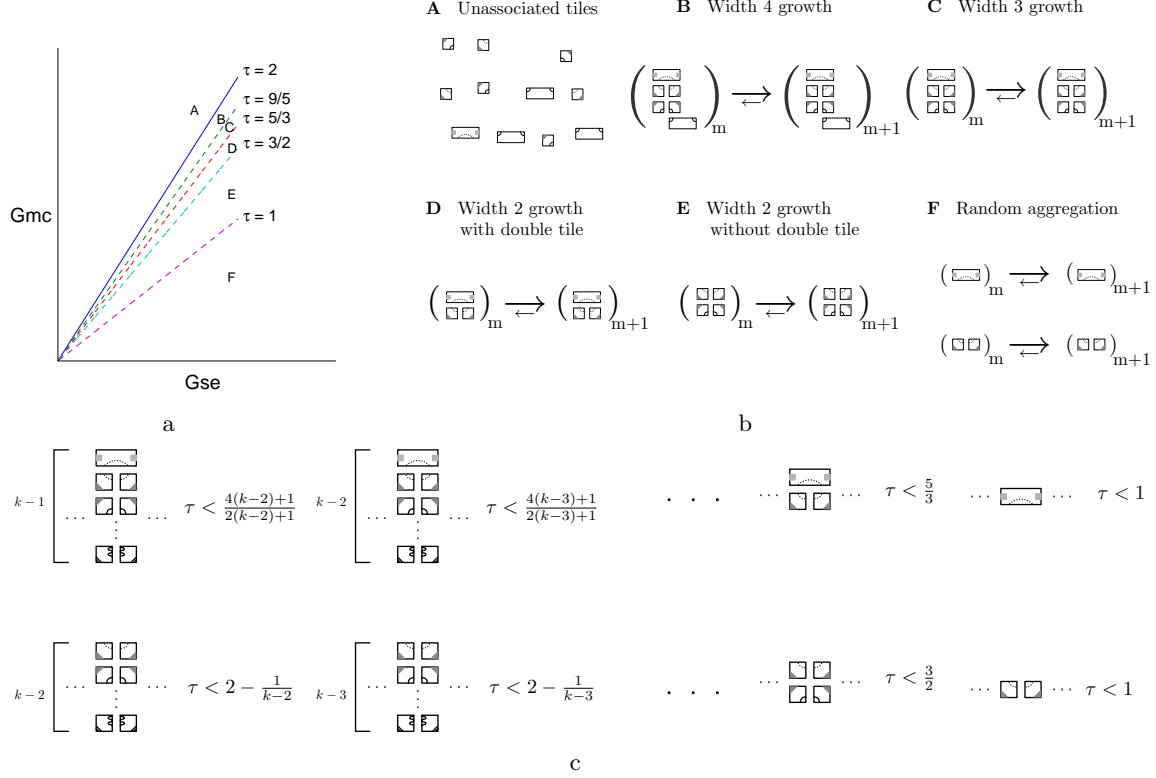


Figure 3.5: **Physical conditions where zig-zag polymer elongation is favorable.** G_{mc} ($\ln(\text{tile concentration})$) and G_{se} (bond strength) define a set of physical conditions for zig-zag tile assembly. $\tau = \frac{G_{mc}}{G_{se}}$. **(a)** Phase diagram of the width 4 zig-zag tile set. In phase A, above the line $\tau = 2$, no assembly reactions are favorable, whereas in regimes B, C, D, E, and F, progressively more types of assemblies (shown in (b)) become favorable. **(b)** The polymeric assemblies which become favorable in the regimes B–F shown in (a). Elongation of a polymer is favorable when the energy change ΔG for adding a repeat unit becomes negative. If adding a repeat unit adds n tiles and creates b new bonds, $\Delta G = nG_{mc} - bG_{se}$. For the polymer shown for regime C, for example, $\Delta G = 5G_{mc} - 9G_{se}$, which is negative when $\tau = \frac{G_{mc}}{G_{se}} < \frac{9}{5}$. Polymers shown for earlier regimes are also favorable in later phases: the polymer shown for regime B is favorable in regimes C–F and so on. **(c)** The assemblies that can form from a zig-zag tile set of width k and the physical conditions (in terms of τ) in which these assemblies becomes favorable. For a zig-zag tile set of width k , only full-width ribbons are favorable when $2 < \tau < \frac{4(k-2)+1}{2(k-2)+1} = 2 - \frac{1}{2k-3}$.

to grow, it is enough to show that for each of these polymer types, longer polymers have a more positive free energy than shorter ones.

To characterize the energy landscape formally, we consider the important classes of polymeric assemblies and evaluate their free energies. Figure 3.5b, B–F show the 6 main types of polymeric assemblies² for ribbons of width 4 by indicating the repeat group that may be added (by a series of

²“Imperfect” long assemblies, for example an assembly with irregular width, can be considered as a member of the smallest polymer class that contains a larger, more complete assembly. Since removing tiles from a “perfect” assembly strictly increases its free energy, these “imperfect” assemblies have strictly lower concentrations than their

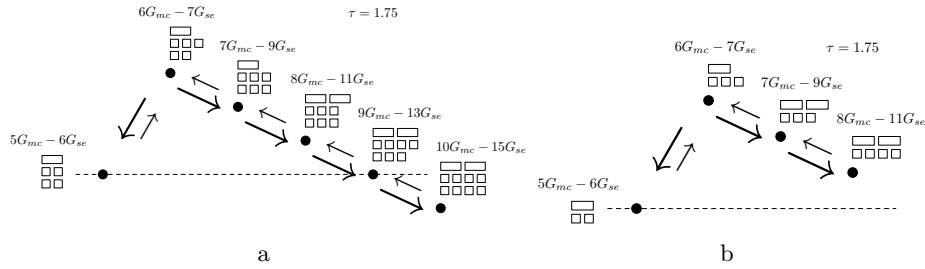


Figure 3.6: **Zig-zag polymerization reactions.** The addition of a polymer unit to a thin assembly consists of an initial unfavorable accretion reaction followed by a series of favorable accretion reactions. **(a)** A favorable polymerization reaction. The positive free energy change from the four favorable accretion reactions is larger than the negative energy change from the initial unfavorable accretion reaction. Thus, polymers of width 3 are favorable where $\tau = 1.75$. **(b)** An unfavorable polymerization reaction. The positive free energy change from the two favorable accretion reactions is not large enough to compensate for the negative energy change from the initial unfavorable accretion reaction, so that polymers of width 2 are unfavorable where $\tau = 1.75$.

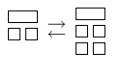
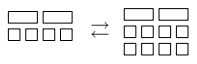
Reaction	ΔG	τ below which reaction becomes favorable
	$3G_{se} - 2G_{mc}$	$\frac{3}{2}$
	$7G_{se} - 4G_{mc}$	$\frac{7}{4}$
$\left(\begin{array}{c} \square \\ \square \end{array}\right)_m \rightleftharpoons \left(\begin{array}{c} \square \\ \square \\ \square \end{array}\right)_m$	$(2m - 1)G_{se} - mG_{mc}$	$\frac{2m-1}{m}$

Figure 3.7: **Physical conditions (in terms of τ) where an increase in assembly width is favorable.** Like the polymerization of thin ribbons, a reaction to produce a wider assembly from a thinner one consists of an initial unfavorable accretion reaction followed by a series of favorable accretion reactions to complete the new row. The number of favorable reactions determines for what τ values the overall reaction is favorable.

accretion reactions, as shown in Figure 3.6) to extend the polymer. To determine whether adding a repeat group results in a higher or lower energy assembly, we evaluate $\Delta G = G(A_{m+1}) - G(A_m) = \Delta n G_{mc} - \Delta b G_{se}$ where A_m is a polymeric assembly with m repeat units. If ΔG is negative, then longer polymeric assemblies of this type are more favorable and we can expect this kind of assembly to grow at some rate. This gives a linear condition on G_{se} and G_{mc} , specifying a regime of physical conditions in which a certain class of long assembly is favorable. For example, for polymer type **E**, each repeat unit adds 4 tiles ($\Delta n = 4$) and 6 bonds ($\Delta b = 6$), so these polymers grow if $4G_{mc} - 6G_{se} < 0$, i.e., $\frac{G_{mc}}{G_{se}} < \frac{3}{2}$. Similar calculations result in the phase diagram shown in Figure 3.5a, which shows the melted phase A, in which no polymers are favorable, and the crystalline phase divided into regimes B–F wherein one additional type of polymer becomes favorable in each successive regime. In all these calculations, the ratio $\tau \stackrel{def}{=} \frac{G_{mc}}{G_{se}}$ plays a critical role.

Figure 3.5c shows the $2k - 3$ classes of polymeric assemblies for the width k zig-zag tile set (excluding the full width ribbons) along with the condition on τ that governs when elongation of the respective polymers is favorable. When $2 > \tau > 2 - \frac{1}{2k-3}$, zig-zag growth is favorable, but the elongation of all less than full-width polymers is unfavorable.

The table in Figure 3.7 enumerates the assemblies for which growing wider (rather than longer) is favorable. The table shows that very long assemblies can favorably grow wider even when τ is close to 2, so for optimal nucleation control it is necessary that lengthening of thin assemblies be unfavorable. Otherwise, a favorable path to nucleation exists: an assembly can grow longer until it is favorable for it to grow wider. The assembly would then grow to full width.

An example of the difference in the energy landscape between the regime where only full width polymers are favorable ($2 < \tau < 2 - \frac{1}{2k-3}$), and a regime where other polymers are favorable can be seen in Figure 3.8. When $2 < \tau < 2 - \frac{1}{2k-3}$, as in the first and third landscapes in this figure, the critical nuclei (denoted by the larger circles) are of width $k - 1$ (or width k) for the two widths shown. The critical nuclei for the tile set of width 8 are more unfavorable than those of width 4. In contrast, when τ is outside this regime, as in the second and fourth landscapes, a wider zig-zag tile set does not change the critical nucleus size and should not cause an exponential decrease in spurious nucleation rates.

corresponding “perfect” assembly.

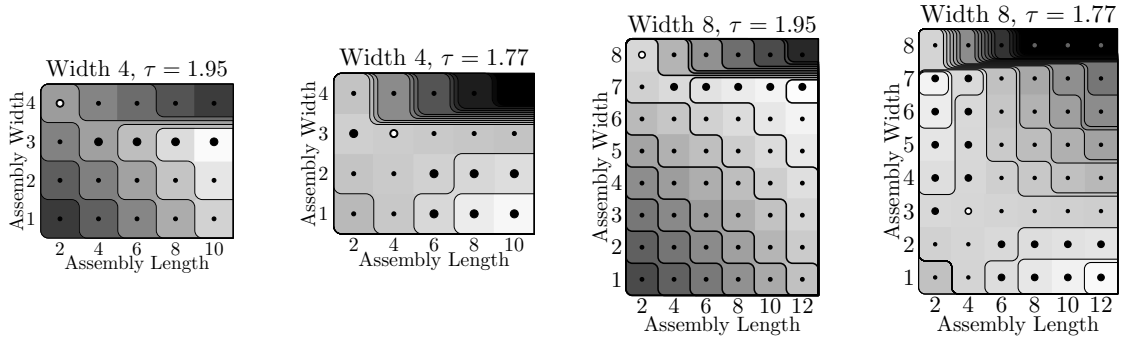


Figure 3.8: **Example energy landscapes.** Coarse-grained depictions of the energy landscapes for two zig-zag tile sets of different widths under two different physical conditions. Each square in the grid is an assembly of one width and length. The shading in the square represents the energy of a rectangular assembly of those dimensions. Darker is more favorable. Contour lines group assemblies of similar energies. Larger circles denote assemblies that are critical nuclei – those assemblies that can both through a series of favorable increases in length or width reach full width and through a series of favorable decreases in length or width melt into single tiles. The most favorable critical nucleus, (the principal critical nucleus) is denoted by a large hollow circle. When $\tau = 1.95$, $2 < \tau < 2 - \frac{1}{2k-3}$ (Figure 3.7(c)) for both $k = 4$ and $k = 8$, so the critical nuclei are of width $k - 1$ or k . Under these conditions, the most favorable path to nucleation for both tile sets is for a crystal of length 2 to grow to full width. Thus, the barrier to nucleation for a tile set of width 8 is higher than the barrier to nucleation for a tile set of width 4. In contrast, when $\tau = 1.77$, the principal critical nucleus is the same for both tile sets: it is an assembly of width 3 and length 4. Under these conditions, the spurious nucleation rate will not be appreciably smaller with the wider zig-zag tile set.

Thus, the regime for optimal control over nucleation is limited to $2 > \tau > 2 - \frac{1}{2k-3}$. The primary theorem of the next section will be relevant only in this region. While the desirable region in the phase diagram appears small, a slow anneal from a high temperature where $\tau \gg 2$ to a temperature in which $\tau < 1$ will pass through this regime, and a slow enough anneal will allow the bulk of the reaction to take place in this regime. Therefore, it is reasonable to consider a mechanism for the control of nucleation which is valid only in this narrow range of physical conditions. In the next section, we analyze the nucleation rates of the zig-zag tile set within this regime.

3.5 An Asymptotic Bound on Spurious Nucleation Rates

The kinetic Tile Assembly Model predicts the concentration of each assembly at all times. For most tile sets, the number of possible assemblies is large, and the individual concentrations of many kinds of intermediate assemblies are not necessarily of interest. Instead, it is often helpful to talk about the concentration of a **class** $\mathcal{C} \subset \mathcal{A}$ of assemblies, $[\mathcal{C}] = \sum_{A \in \mathcal{C}} [A]$.

The derivative of the concentration of a class of assemblies, $\frac{d[\mathcal{C}]}{ds} = \sum_{A \in \mathcal{C}} \frac{d[A]}{ds}$, can be calculated as the difference between the rate at which assemblies join the class and that at which they leave the class. Reactions which produce new members of the class from assemblies not in the class are the **inward perimeter reactions**, $R^{in} = \{A+t \rightarrow B+t, A \rightarrow B, t_1+t_2 \rightarrow B+t_1+t_2 : A \notin \mathcal{C}, B \in \mathcal{C}\}$. Reactions which use up members of the class to produce assemblies not in the class (or single tiles) are the **outward perimeter reactions** $= R^{out} = \{B+t \rightarrow A+t, B \rightarrow A, B \rightarrow \emptyset : A \notin \mathcal{C}, B \in \mathcal{C}\}$.

Define the **flux** across a set of reactions R at time s as

$$F(R, s) = \sum_{A+t \rightarrow B+t \in R} k_f [A] e^{-G_{mc}} + \sum_{B \rightarrow A \in R} k_f e^{G^\circ(B) - G^\circ(A)} [B] + \sum_{t_1+t_2 \rightarrow A+t_1+t_2 \in R} k_f e^{-2G_{mc}} + \sum_{A \rightarrow \emptyset \in R} k_f e^{G^\circ(A)} [A] \quad (3.2)$$

Then $\frac{d[\mathcal{C}]}{ds}(s) = F(R^{in}, s) - F(R^{out}, s)$.

We will use these formalisms to bound the rate of spurious nucleation in a zig-zag tile set of width k . The **spuriously nucleated assemblies** of width k will be denoted \mathcal{C}_k . Let the top tile in Figure 3.3a be designated \mathbf{t}_t the bottom tile be designated \mathbf{t}_b , and the seed tile be designated \mathbf{t}_s . Formally,

$$\mathcal{C}_k = \{A \in \mathcal{A} : \exists (x, y), (w, z) \in \mathbb{Z}^2 \text{ s.t. } A(x, y) = \mathbf{t}_t, A(w, z) = \mathbf{t}_b, \text{ and } \forall (q, r) \in \mathbb{Z}^2, A(q, r) \neq \mathbf{t}_s\} \quad (3.3)$$

Note that the assemblies in \mathcal{C}_k do not contain a seed tile because we are measuring the rate of formation if zig-zag ribbons without seed tiles.

The inward perimeter reactions for $[\mathcal{C}_k]$, which we call the **spurious nucleation reactions** and denote by R_k^{in} , are the reactions for which the product is a full width assembly, but the reactant is not. In other words, they are the addition reactions which produce width k assemblies from assemblies of width $k-1$ by adding either a top or a bottom double tile (Figure 3.9). As shown in Section 3.4, when $2 < \frac{G_{mc}}{G_{se}} < \frac{1}{2k-3}$, these reactions demarcate the point at which sustained

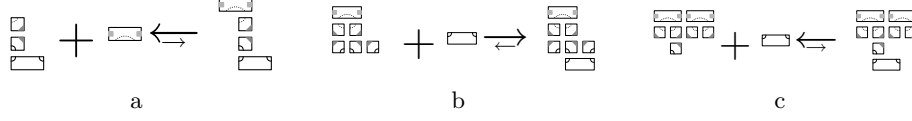


Figure 3.9: **Sample spurious nucleation reactions.** Three spurious nucleation reactions for a zig-zag tile set of width 4. The reaction may be either favorable or unfavorable. In (b), the addition is favorable when $G_{mc} \approx 2G_{se}$, because two new bonds are formed; In (a) and (c), the addition is unfavorable under these conditions because in each reaction only one new bond is formed.

growth can proceed by exclusively favorable steps. The outward perimeter reactions, which we call the **ribbon shrinking reactions** and denote by R_k^{out} , are those in which a tile falls off a full width assembly to produce an assembly of width $k - 1$. For assemblies that have suffered a ribbon shrinking reaction, there is also a downhill path to complete melting in the energy landscape of the type shown in Fig. 3.8 when $2 < \frac{G_{mc}}{G_{se}} < \frac{1}{2k-3}$.

The overall rate of spurious nucleation of width k zig-zag crystals (in units of Molar per second),

$$n_k(s) = \frac{d[C_k]}{ds}(s) = F(R_k^{in}, s) - F(R_k^{out}, s),$$

may be integrated over time to obtain the total concentration of spuriously nucleated assemblies. Furthermore, an upper bound on $n_k(s)$ similarly translates into an upper bound on the concentration of spuriously nucleated assemblies. Because the growth path for full-width ribbons is so favorable (zig-zag growth), one such bound is obtained by neglecting the ribbon shrinking reactions and considering just the spurious nucleation reactions:

$$n_k^+(s) = F(R_k^{in}, s) > n_k(s).$$

Theorem 1 For a zig-zag tile set of width $k > 2$, if $2 > \frac{G_{mc}}{G_{se}} > 2 - \delta$, $\delta < \frac{1}{2k-3}$, and $G_{se} > \frac{2k \ln 2}{1 - (2k-3)\delta}$, then for all times s , $n_k(s) < 4k_f e^{(\delta-k)G_{se}}$.

Proof Since $n_k^+(s) < n_k(s)$, we can prove the theorem by showing that $n_k^+(s) < 4k_f e^{(\delta-k)G_{se}}$. All the spurious nucleation reactions are addition reactions, so if we compute $n_k^+(s)$ using Equation 3.2, the second and fourth terms of the expression are both zero. Spuriously nucleated assemblies are defined as assemblies of width k , so the reactants in the spurious nucleation reactions are of width $k - 1$. (Only accretion reactions are allowed.) For a tile set of width $k > 2$, the third term of Equation 3.2—the contribution of the interaction of two tiles—also drops out. Therefore, for a tile set of width $k > 2$ with spurious nucleation reactions R_k^{in} ,

$$n_k(s) \leq n_k^+(s) = \sum_{A+t \rightarrow B+t \in R_k^{in}} k_f [A] e^{-G_{mc}}, \quad (3.4)$$

where $[A]$ is the concentration of assembly A at time point s .

While it is in general difficult to calculate $[A]$ at an arbitrary time point, the following lemma shows that the concentration of an assembly can be bounded by its concentration at steady state,

which is easy to compute:

Lemma 3 *In a mass-action powered accretion model of self-assembly that has an initial state containing only single tiles, every assembly has a concentration less than or equal to its steady-state concentration³.*

Proof Suppose that this lemma is not true. Then there is a time at which the concentrations of one or more assemblies exceed their values at steady state. Since the concentrations of all assemblies are zero initially, there must be a first time s at which for at least one assembly A , $[A] = [A]_{ss}$. At this time, the concentrations of all other assemblies are either at or below their respective steady-state concentrations. From Section 3.3, the rate of change of $[A]$ is given by the formula

$$\frac{d[A]}{ds} = k_f \left(\sum_{\substack{A+t \rightarrow B+t, \\ B \rightarrow A \in R}} e^{G^\circ(B) - G^\circ(A)} [B] - [A] e^{-G_{mc}} + \sum_{\substack{B+t \rightarrow A+t, \\ A \rightarrow B \in R}} [B] e^{-G_{mc}} - e^{G^\circ(A) - G^\circ(B)} [A] + \sum_{\substack{t_1+t_2 \rightarrow A+t_1+t_2, \\ A \rightarrow \emptyset \in R}} e^{-2G_{mc}} - e^{G^\circ(A)} [A] \right).$$

Consider a single term in the second summation, $[B] e^{-G_{mc}} - e^{G^\circ(A) - G^\circ(B)} [A]$, involving some assembly B . We know that $[A]$ has reached its steady-state concentration, so $[A] = e^{-G(A)}$. By assumption, $[B] \leq [B]_{ss} = e^{-G(B)}$. Assembly A includes one more tile, t , than does assembly B , so $G^\circ(A) - G^\circ(B) = G(A) - G(B) - G_{mc}$. Therefore,

$$\begin{aligned} [B] e^{-G_{mc}} - e^{G^\circ(A) - G^\circ(B)} [A] &= [B] e^{-G_{mc}} - e^{G(A) - G(B) - G_{mc}} [A] \\ &= [B] e^{-G_{mc}} - e^{G(A) - G(B) - G_{mc}} e^{-G(A)} \\ &= [B] e^{-G_{mc}} - e^{-G(B)} e^{-G_{mc}} \\ &= e^{-G_{mc}} \left([B] - e^{-G(B)} \right) \\ &\leq 0. \end{aligned}$$

Similarly, for an assembly B that is a term in the first summation, B has the extra tile t so that $G^\circ(B) - G^\circ(A) = G(B) - G(A) - G_{mc}$. The term can be simplified to

$$\begin{aligned} e^{G^\circ(B) - G^\circ(A)} [B] - [A] e^{-G_{mc}} &= e^{G(B) - G(A) - G_{mc}} [B] - e^{-G(A)} e^{-G_{mc}} \\ &\leq e^{G(B) - G(A) - G_{mc}} e^{-G(B)} - e^{-G(A)} e^{-G_{mc}} \\ &= 0. \end{aligned}$$

The terms in the third summation are also non-positive, since

³The concentration of the class \mathcal{C}_k , which at the conditions we consider contains an infinite number of assemblies, is actually infinite at steady state. The inward flux, as we will show, is finite because the concentration of unnuclated assemblies stays finite at steady state, even though there are also an infinite number of unnuclated assemblies.



Figure 3.10: **Assembly dimensions of rectangular assemblies.** (a) A $k - 1 = 3$ by $l = 8$ assembly. (b) A $k - 1 = 3$ by $l = 7$ assembly

$$\begin{aligned}
 e^{-2G_{mc}} - e^{G^\circ(A)}[A] &= e^{-2G_{mc}} - e^{G^\circ(A)}e^{-G(A)} \\
 &= e^{-2G_{mc}} - e^{G(A)-2G_{mc}}e^{-G(A)} \\
 &= 0.
 \end{aligned}$$

The change in concentration $\frac{d[A]}{ds}(s)$ is composed entirely of terms of this form. Since each of these terms is non-positive, $\frac{d[A]}{ds}(s)$ is non-positive when $[A] = [A]_{ss}$ and $[A]$ can never rise above its steady-state value. Thus, $[A] \leq [A]_{ss}$.

As in Lemmas 1 and 2, this proof also applies to a model of self-assembly with arbitrary stoichiometry and sticky end strengths. (Explicit proof is not shown.)

Lemma 3 implies that

$$F(R_k^{in}, s) \leq \sum_{A+t \rightarrow B+t \in R_k^{in}} k_f [A]_{ss} e^{-G_{mc}}$$

where $[A]_{ss}$ is the concentration of assembly A at steady state.

Partitioning the summation according to the length of the reactant assembly gives

$$F(R_k^{in}, s) \leq \sum_{l=1}^{\infty} \sum_{\substack{\text{length}(A)=l \\ A+t \rightarrow B+t \in R_k^{in}}} k_f [A]_{ss} e^{-G_{mc}}. \quad (3.5)$$

To be a reactant in a spurious nucleation reaction, A must have a width of $k-1$. By assumption, A cannot have any mismatches⁴. Thus, each assembly A in the preceding summation can be viewed as a $k-1$ by l rectangular assembly of the type shown in Figure 3.10 with zero or more tiles missing⁵. $2G_{se} > G_{mc}$ by assumption, so the free energy of a $k-1$ by l assembly cannot be more favorable than the free energy of the $k-1$ by l rectangle that contains it, since any missing tiles in the rectangle could be added by a series of favorable reactions. Therefore, the concentration of any $k-1$ by l assembly at steady-state must be no larger than the concentration of its corresponding $k-1$ by l rectangular assembly. Note that this bound is very loose, since most assembly types have several tiles attached by only one bond and therefore have a higher free energy. Let $A_{k-1,l}$ be a $k-1$ by l rectangular assembly, and $C(k-1, l)$ be the number of assemblies of width $k-1$ and

⁴Because all bonds are unique, a potential tile addition to any assembly either matches the assembly on all sides, such that no errors occur, or matches on no sides, such that the addition does not produce a connected assembly.

⁵It could also be a subset of a rectangular assembly with top instead of bottom tiles, but the free energy of both kinds of assemblies is the same. To account for this, we include a 2 pre-factor in Equation 3.6, thereby writing counting both rectangles with top tiles and rectangles with bottom tiles.

length l . Each assembly can bind a single tile in up to l locations along either the top or bottom edge. Thus,

$$\begin{aligned} F(R_k^{in}, s) &< \sum_{l=1}^{\infty} \sum_{\substack{\text{length}(A)=l \\ A+t \rightarrow B+t \in R_k^{in}}} k_f [A_{k-1,l}]_{ss} e^{-G_{mc}} \\ &\leq \sum_{l=1}^{\infty} C(k-1, l) l k_f [A_{k-1,l}]_{ss} e^{-G_{mc}}. \end{aligned}$$

A counting argument shows that $C(k-1, l) < 2^{(k-1)l+1}$, so

$$F(R_k^{in}, s) < 2 \sum_{l=1}^{\infty} 2^{(k-1)l} l k_f [A_{k-1,l}]_{ss} e^{-G_{mc}}. \quad (3.6)$$

The steady-state concentration of an unseeded assembly with n tiles and b bonds is given by $[A]_{ss} = e^{-nG_{mc}+bG_{se}}$. The assembly $A_{k-1,l}$ contains $(k-2)l$ small tiles and $\lceil l/2 \rceil$ top (or bottom) tiles. There are $(l-1)(k-2)$ horizontal bonds between small tiles and $\lceil l/2 \rceil - 1$ horizontal bonds between large tiles. In addition, there are up to l vertical bonds in each of the $k-2$ spaces between rows of tiles. Therefore,

$$[A_{k-1,l}]_{ss} \leq \exp(-((k-2)l + l/2)G_{mc} + ((k-2)(l-1) + l/2 + (k-2)l)G_{se}).$$

Applying the assumption $G_{mc} > (2-\delta)G_{se}$ and simplifying,

$$[A_{k-1,l}]_{ss} < \exp\left[(2-k)G_{se} + \left(k\delta - \frac{1}{2} - \frac{3\delta}{2}\right)lG_{se}\right].$$

Thus,

$$F(R_k^{in}, s) < 2k_f e^{-G_{mc}} e^{(2-k)G_{se}} \sum_{l=1}^{\infty} l 2^{(k-1)l} e^{(k\delta - \frac{1}{2} - \frac{3\delta}{2})lG_{se}}.$$

Since $k\delta - \frac{1}{2} - \frac{3\delta}{2} < 0$ when $\delta < \frac{1}{2k-3}$, bounding G_{se} from below preserves the inequality. Therefore, when $G_{se} > \frac{2k \ln(2)}{1-(2k-3)\delta}$,

$$\begin{aligned} F(R_k^{in}, s) &< 2k_f e^{-G_{mc}} e^{(2-k)G_{se}} \sum_{l=1}^{\infty} l 2^{(k-1)l} e^{(k\delta - \frac{1}{2} - \frac{3\delta}{2})l \frac{2k \ln(2)}{1-(2k-3)\delta}} \\ &= 2k_f e^{-G_{mc}} e^{(2-k)G_{se}} \sum_{l=1}^{\infty} l 2^{(k-1)l} e^{-kl \ln 2} \\ &= 2k_f e^{-G_{mc}} e^{(2-k)G_{se}} \sum_{l=1}^{\infty} l 2^{-l} \\ &= 4k_f e^{-G_{mc}} e^{(2-k)G_{se}} \\ &< 4k_f e^{(\delta-k)G_{se}}. \end{aligned}$$

This theorem says that the spurious nucleation rate, n_k , decreases exponentially with k and with G_{se} , within the limits of applicability of the theorem—which requires larger G_{mc} for larger k , and hence slower growth rates. The strength of the theorem, therefore, lies in the extent to which spurious nucleation decreases *faster* than the growth rate (rows added per unit time), r_k , of seeded crystals. These relative rates translate into the degree of purity that can be obtained when attempting to grow seeded crystals: suppose the concentration of seeds is c , and they are grown to a length L during a time period $s = L/r_k$. The concentration of unseeded crystals that will have spuriously nucleated in that time is $s \cdot n_k = L \cdot \frac{n_k}{r_k}$, i.e., the ration of crystals that were spuriously nucleated to the concentration of seeds is $\frac{L}{c} \cdot \frac{n_k}{r_k}$. (When we use n_k without specifying a particular time, we mean the asymptotic value, which is an upper bound.) Regardless of what length or amount of seeded crystals is desired, reducing $\frac{n_k}{r_k}$ is the relevant metric for increasing the yield of desired structures.

One way to study the trade-off between n_k and r_k is to ask, given a target growth rate r , what is the lowest nucleation rate that can be achieved by adjusting G_{mc} and G_{se} while maintaining $r_k = r$? Previous work [Win98] has shown that near the $\tau = 2$ phase boundary that is relevant to our theorem, the growth rate is closely approximated by

$$r_k = \frac{k_f}{k-1} (e^{-G_{mc}} - e^{-2G_{se}}),$$

measured in layers per second. The lowest nucleation rate for a given target growth rate r is then

$$n_k^*(r) = \min_{\substack{G_{se}, G_{mc} \\ \text{s.t. } r_k=r}} n_k.$$

A plot of $n_k^*(r)$ vs. r , if it could be calculated, would reveal how much the spurious nucleation rate can be decreased for a given decrease in the growth rate, for zig-zag crystals of a given width. Theorem 1 only gives us an upper bound on $n_k^*(r)$, but even so, this already gives us a characterization of the advantage provided by wider zig-zag crystals.

Specifically, choosing $2G_{se} - G_{mc} = \epsilon = \ln k$, $\delta = \frac{1}{2} \left(\frac{1}{2k-3} \right)$ and $G_{se} > 4k \ln k$, Theorem 1 guarantees that $n_k^* < n_k < 4ek_f e^{-kG_{se}}$, while $r_k = \frac{k_f}{k-1} (e^{\epsilon-2G_{se}} - e^{-2G_{se}}) = k_f e^{-2G_{se}}$. Thus, under these conditions, the ratio of $\frac{n_k}{r_k}$ also decreases exponentially, suggesting that seeded zig-zag crystals can be grown with exponentially greater yield as width increases.

3.6 Numerical Estimates of Spurious Nucleation Rates

Having shown in the previous section that zig-zag tile sets can be designed to achieve arbitrarily low spurious nucleation rates relative to the growth rates, we now ask whether the nucleation barrier provided by zig-zag tile sets is sufficient for practical implementation in the laboratory. There are two main concerns: first, as each tile must be synthesized, k must be small (6 is currently practical, while 50 is currently too large); second, assembly time must not be too long. (Growing 1000 layers of seeded crystals with less than 1% spurious nucleation—which we refer to as the “typical reaction”—seems like a reasonable goal to accomplish within one week.) As the asymptotic bounds of Section 3.5 are too loose for obtaining a realistic evaluation of nucleation rates for small k , we now develop more accurate numerical calculations and stochastic simulations for estimating

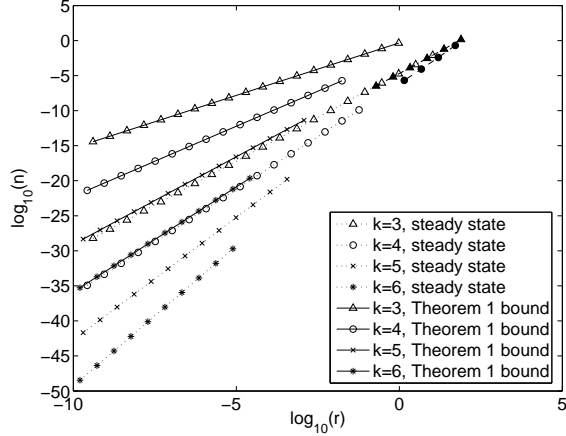


Figure 3.11: **Calculated steady-state nucleation rates.** Asymptotic bound on the nucleation rate ($n_k = 4k_f e^{(\delta-k)G_{se}}$) given by Theorem 1 and numerical calculations of nucleation rates at steady state as described in Section 3.6.1. The graph compares the growth rate r_k (in layers/s) and the rate of spurious nucleation events, n_k^+ (in M/s), for $2G_{se} - G_{mc} = \epsilon = 0.1$. Since the forward rate constant k_f has not been measured experimentally for tile-based assembly, we use $k_f = 6 \times 10^5$ /M/s based on typical oligonucleotide hybridization rates [QW89]. Filled triangles and circles denote rates of spurious nucleation events for $k = 3$ and $k = 4$ respectively, estimated from stochastic simulations, and are plotted in more detail in Figure 3.12. They are plotted here to illustrate that these measured rates are consistent with the calculated rate of steady-state spurious nucleation.

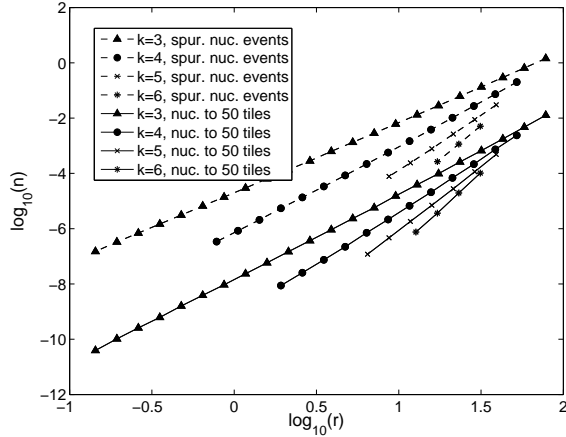


Figure 3.12: **Estimates of nucleation rates from simulations.** Estimates of the ratio between assembly speed r_k and the average, over the time needed for seeded crystals to grow 1000 layers, of the rate of spurious nucleation events ($\frac{1}{s} \int_0^s n_k^+(s)$) or the overall spurious nucleation rate ($\frac{1}{s} \int_0^s n_k(s)$), as measured by stochastic simulations either (a) counting the frequency of reactions that create full-width ribbons, i.e., R_k^{in} , or (b) counting the number of assemblies that have 50 tiles or more. $\epsilon = 0.1$. Simulations are only practical for high concentrations, and the numerical calculations of Fig. 3.11 are only provably valid for lower concentrations.

spurious nucleation rates.

The analysis in Section 3.5 overestimates the spurious nucleation rate in three ways. First, it overestimates the concentration of almost all kinds of assemblies by assuming they have the same concentration as a rectangular assembly of the same length and width, and it overcounts the number of different types of assemblies. Second, Lemma 3 shows that the spurious nucleation rate at steady state is the maximal spurious nucleation rate. However, it may take longer to approach steady state than the time needed to run a “typical reaction,” and far from steady state, the spurious nucleation rate may be much smaller than the rate at steady state. Lastly, this analysis defines a spurious nucleation event for a zig-zag tile set of width k as a reaction that produces an assembly of width k , and neglects the backward reaction. In practice, many reactions that form an assembly of width k are unfavorable, so that the product assembly frequently shrinks back to a sub-critical size instead of growing larger. Furthermore, when conditions only slightly favor growth, even assemblies containing several layers have a reasonable chance of shrinking to nothing before they grow substantially. I.e., we expect n_k to be significantly smaller than n_k^+ in this case.

While it is not possible to compute the nucleation rate exactly, in this section, we describe three numerical techniques that correct each inaccuracy for zig-zag tile sets of widths $k = 3, 4, 5,$ and 6 . In Section 3.6.1, we much more accurately compute the rate at which ribbons of width k are formed at steady state. These computations show that the asymptotic bound of Theorem 1 is too high by at least 4 orders of magnitude for the range of parameters studied. In Section 3.6.2 we use a stochastic simulation of tile assembly to estimate the rate of spurious nucleation reactions R_k^{in} . Our results indicate that spurious nucleation reactions occur during a “typical reaction” at virtually the same rate as at steady state. In Section 3.6.3, we use the stochastic simulation to investigate whether the rate of spurious nucleation reactions ($n_k^+ = F(R_k^{in}, s)$) in a typical reaction accurately predicts the rate at which large assemblies appear (which at steady state is equivalent to $n_k(s) = F(R_k^{in}, s) - F(R_k^{out}, s)$). We find that for the range of parameters studied, at least 99% of assemblies that reach full width will melt before growing into large crystals, and thus our other estimates of spurious nucleation rates may be overestimates of n_k by at least this factor. In Section 3.6.4, we show that these results together indicate that a zig-zag tile set of width 5 or 6 should be large enough to prevent almost all spurious nucleation in a “typical reaction”, while maintaining reasonable assembly speeds. We conclude with an important caveat to these results. Our results are derived under a kTAM model where assemblies may grow only through the addition of single tiles. In contrast, in experiments small assemblies may aggregate rather than growing exclusively by single tile additions, thus potentially producing nuclei that reach a critical size more quickly than our simulations indicate.

3.6.1 Spurious Nucleation Rates at Steady State

Recall that for a zig-zag tile set of width $k > 2$, the steady-state rate of spurious nucleation reactions is given by the sum

$$n_k^+ = \lim_{s \rightarrow \infty} F(R_k^{in}, s) = \sum_{l=1}^{\infty} \sum_{\substack{A+t \rightarrow B+t \in R_k^{in} \\ \text{s.t. } \text{length}(A)=l}} k_f[A]_{ss} e^{-G_{mc}},$$

which ignores the rate at which spuriously nucleated assemblies dissolve back into pre-nucleated assemblies. While $[A]_{ss}$ is known (if A has n tiles and b bonds, $[A]_{ss} = e^{bG_{se} - nG_{mc}}$), it is not

practical to compute the sum exactly because there are an infinite number of spurious nucleation reactions. Additionally, it can be impractical to evaluate the inner sum even for a single value of l : no efficient algorithm is known (see e.g. [Gol94] for the related problem of counting polyominoes) for exactly enumerating the reactions in R_k^{in} . The number of distinct reactions increases exponentially with the length of A , so that it is prohibitive to calculate all but the first terms of the sum.

Despite these difficulties, the expression can be calculated with small, known error bounds for many k . The following lemma shows that under many reaction conditions of interest, the sum converges quickly, so that its value can be approximated by summing only the first few terms:

Lemma 4 *When $G_{se} > (\ln 10)(k - 2) + \ln 4$, $G_{mc} = 2G_{se} - \epsilon$, $0 \leq \epsilon < \frac{1}{2k-3}$ and l is even,*

$$\sum_{p=l+1}^{\infty} \left(\sum_{\substack{A+t \rightarrow B+t \in R_k^{in} \\ \text{s.t. } \text{length}(A)=p}} k_f[A]_{ss} e^{-G_{mc}} \right) < 2 \left(\sum_{\substack{A+t \rightarrow B+t \in R_k^{in} \\ \text{s.t. } \text{length}(A)=l}} k_f[A]_{ss} e^{-G_{mc}} \right)$$

Proof We start by re-writing the lemma to use convenient notation to refer to the inner sums within the series for n_k^+ , which refer to the rate of spurious nucleation events involving assemblies A of width $k - 1$ and length l :

$$N_p = \sum_{\substack{A+t \rightarrow B+t \in R_k^{in} \\ \text{s.t. } \text{length}(A)=p}} k_f[A]_{ss} e^{-G_{mc}},$$

such that $n_k^+ = \sum_{l=1}^{\infty} N_l$. Now, Lemma 4 may be stated as:

When $G_{se} > (\ln 10)(k-2) + \ln 4$, $G_{mc} = 2G_{se} - \epsilon$, $0 \leq \epsilon < \frac{1}{2k-3}$ and l is even, then $\sum_{p=l+1}^{\infty} N_p < 2N_l$.

To prove this lemma, we will prove two sub-lemmas. First,

Lemma 5 *If $G_{se} > (\ln 4)(k - 2) + \ln \frac{12}{5}$, $G_{mc} = 2G_{se} - \epsilon$, l is even and $0 \leq \epsilon < \frac{1}{2k-3}$, then $N_{l+1} < \frac{1}{2}N_l$.*

Proof We will partition the assemblies of length $l + 1$ into classes corresponding to assemblies of length l . We will then show the total spurious nucleation rate of reactions containing the assemblies in each class is at least twice as small as the spurious nucleation rate of reactions containing their corresponding assembly. The class of assemblies of length $l + 1$ corresponding to an assembly B of length l will be denoted \hat{B} .

To assign the assemblies to classes, we introduce a procedure that takes an assembly A of width $k - 1$ and length $l + 1$, and then ‘‘condenses’’ its right end to yield an assembly B with width $k - 1$ and length l . Specifically, A and B are identical except for the last two columns of A and the last column of B ; there, if A had a tile in either the ultimate or penultimate column in some particular row, then B will have a tile in its last column in the same row. Recall that for valid zig-zag assemblies,

if a tile is present in a particular spot, its tile type is determined by its neighbors – thus, we don't have to specify tile types in our condensation procedure, since there is no choice. Formally, we say that $B = \mathbf{condensation}(A)$ if for all $0 \leq a < k - 1$, $0 \leq b < l - 1$: $\tilde{A}(a, b) = \tilde{B}(a, b)$, and for all $0 \leq a < k - 1$: $\tilde{B}(a, l - 1) = 0$ iff $\tilde{A}(a, l - 1) = \tilde{A}(a, l) = 0$. Recall that \tilde{A} , the canonical representation of A , begins indexing sites at 0, so the first column has index 0 and the last ($l + 1^{\text{st}}$) column has index l . Also note that since l is even, A cannot have a double tile extending into its last column, so no double tiles are condensed.

To see that for every assembly A , $\mathbf{condensation}(A)$ is connected (and therefore a valid assembly), note first that A is an assembly, so it is connected. Furthermore, the connectivity graph of $B = \mathbf{condensation}(A)$ (with a vertex for each tile and an edge for each abutting pair) is just a graph-theoretic contraction of the connectivity graph of A that combines any two vertices in the same row of the last two columns of A (then possibly adding some extra edges). Therefore, B remains connected. Thus, each A of width $l + 1$ is assigned to a unique, valid assembly B of width l .

Condensation is many-to-one, so there are many assemblies A that condense onto the same smaller assembly B . We assign A to the class corresponding to the assembly $\mathbf{condensation}(A)$, i.e., the class

$$\hat{B} = \{A : \mathbf{condensation}(A) = B\}.$$

For a given assembly B of length l , the elements of \hat{B} , all of length $l + 1$, can be created by adding p tiles ($1 \leq p \leq k - 2$) to the $l + 1^{\text{st}}$ column of B , and then removing h tiles ($0 \leq h \leq p - 1$) from the l^{th} column.

Imagine making these changes one at a time, say from top to bottom, in each row either moving or adding a tile. For each of the $p - h$ tiles that are added to the $l + 1^{\text{st}}$ column where the corresponding tiles in the l^{th} column are not removed, $p - h$ tiles are added to the assembly and no more than $2(p - h) - 1$ bonds may be formed. For the h tiles that are moved from the l^{th} to the $l + 1^{\text{st}}$ column, no tiles are added, and no more bonds can be created (some might even be lost). Therefore, for each such assembly A ,

$$[A]_{ss} \leq e^{-(p-h)G_{mc}} e^{(2(p-h)-1)G_{se}} [B]_{ss}.$$

Let l_A be the number of spurious nucleation reactions that an assembly A is a reactant of. The rate of spurious nucleation events involving assemblies of length $l + 1$ is therefore given by:

$$N_{l+1} = \sum_{\substack{A, C \in \mathcal{A} \\ \text{s.t. } A+t \rightarrow C+t \in R_k^{\text{in}} \\ \text{length}(A)=l+1}} k_f [A]_{ss} e^{-G_{mc}}.$$

We now partition this sum by summing over all smaller assemblies B , and then for each $A \in \hat{B}$

(recall $\hat{B} = \{A \text{ s.t. } \mathbf{condensation}(A) = B\}$) we count the spurious nucleation reactions:

$$\begin{aligned}
&= \sum_{\substack{B \in \mathcal{A} \\ \text{s.t. } \text{length}(B)=l}} \sum_{\substack{A \in \hat{B}, C \in \mathcal{A} \\ \text{s.t. } A+t \rightarrow C+t \in R_k^{\text{in}}}} k_f[A]_{ss} e^{-G_{mc}} \\
&\leq \sum_{\substack{A, B \in \mathcal{A} \\ \text{s.t. } \mathbf{condensation}(A)=B \\ \text{length}(B)=l}} l_A k_f[A]_{ss} e^{-G_{mc}}.
\end{aligned}$$

Partitioning \hat{B} according to the number of tiles added and moved, and using our inequality for $[A]_{ss}$ in terms of $[B]_{ss}$, we have:

$$\leq \sum_{\substack{B \in \mathcal{A} \\ \text{s.t. } \text{length}(B)=l}} \sum_{p=1}^{k-2} \binom{k-2}{p} \sum_{h=0}^{p-1} \binom{p-1}{h} l_A k_f[B]_{ss} e^{-(p-h)G_{mc}} e^{(2(p-h)-1)G_{se}} e^{-G_{mc}}.$$

Under the conditions of the lemma, $G_{mc} > 2G_{se} - \frac{1}{2k-3}$ so that

$$\begin{aligned}
&< \sum_{\substack{B \text{ s.t.} \\ \text{length}(B)=l}} \sum_{p=1}^{k-2} \binom{k-2}{p} \sum_{h=0}^{p-1} \binom{p-1}{h} l_A k_f[B]_{ss} e^{-2(p-h)G_{se}} e^{\frac{p-h}{2k-3}} e^{(2(p-h)-1)G_{se}} e^{-G_{mc}} \\
&= \sum_{\substack{B \text{ s.t.} \\ \text{length}(B)=l}} \sum_{p=1}^{k-2} \binom{k-2}{p} \sum_{h=0}^{p-1} \binom{p-1}{h} l_A k_f[B]_{ss} e^{\frac{(p-h)}{2k-3}} e^{-G_{se}} e^{-G_{mc}} \\
&= \sum_{\substack{B \text{ s.t.} \\ \text{length}(B)=l}} l_A k_f[B]_{ss} \sum_{p=1}^{k-2} \binom{k-2}{p} e^{\frac{p}{2k-3}} e^{-G_{se}} \sum_{h=0}^{p-1} \binom{p-1}{h} e^{\frac{-h}{2k-3}} e^{-G_{mc}}.
\end{aligned}$$

Noting that the inner sums are binomial expansions of (e.g. $(1+x)^n = \sum_{i=0}^n \binom{n}{i} x^i$) or portions thereof, we can simplify further:

$$= \sum_{\substack{B \text{ s.t.} \\ \text{length}(B)=l}} l_A k_f[B]_{ss} \sum_{p=1}^{k-2} \binom{k-2}{p} e^{\frac{p}{2k-3}} e^{-G_{se}} (1 + e^{\frac{-1}{2k-3}})^{p-1} e^{-G_{mc}}.$$

Since for $k > 2$, $\frac{1}{2} < (1 + e^{\frac{-1}{2k-3}})^{-1} < \frac{3}{5}$,

$$\begin{aligned}
&< \sum_{\substack{B \text{ s.t.} \\ \text{length}(B)=l}} \frac{3}{5} l_A k_f[B]_{ss} \sum_{p=1}^{k-2} \binom{k-2}{p} e^{\frac{p}{2k-3}} e^{-G_{se}} (1 + e^{\frac{-1}{2k-3}})^p e^{-G_{mc}} \\
&< \sum_{\substack{B \text{ s.t.} \\ \text{length}(B)=l}} \frac{3}{5} l_A k_f[B]_{ss} \sum_{p=1}^{k-2} \binom{k-2}{p} e^{\frac{p}{2k-3}} 2^p e^{-G_{se}} e^{-G_{mc}} \\
&< \sum_{\substack{B \text{ s.t.} \\ \text{length}(B)=l}} \frac{3}{5} l_A k_f[B]_{ss} \left(1 + 2e^{\frac{1}{2k-3}}\right)^{k-2} e^{-G_{se}} e^{-G_{mc}}.
\end{aligned}$$

Similarly, for $k > 2$, $(1 + 2e^{\frac{1}{2k-3}}) < 4$, and $l_A < l_B + 1$ since the longer assembly A can have at most one more spurious nucleation reaction than B , so

$$\begin{aligned}
&< \sum_{\substack{B \text{ s.t.} \\ \text{length}(B)=l}} \frac{3}{5} (l_B + 1) k_f[B]_{ss} e^{\ln(4)(k-2)} e^{-G_{se}} e^{-G_{mc}} \\
&\leq \sum_{\substack{B \text{ s.t.} \\ \text{length}(B)=l}} \frac{6}{5} l_B k_f[B]_{ss} e^{\ln(4)(k-2)} e^{-G_{se}} e^{-G_{mc}}.
\end{aligned}$$

When $G_{se} > \ln(4)(k-2) + \ln(\frac{12}{5})$,

$$\begin{aligned}
&< \sum_{\substack{B \text{ s.t.} \\ \text{length}(B)=l}} \frac{1}{2} l_B k_f[B]_{ss} e^{-G_{mc}} \\
&= \frac{1}{2} \sum_{\substack{A+t \rightarrow B+t \in R_k^{in} \\ \text{s.t. length}(A)=l}} k_f[A]_{ss} e^{-G_{mc}} = \frac{1}{2} N_l.
\end{aligned}$$

The above sub-lemma takes care of the smaller odd terms, but to show that the entire summation is bounded, we show that the smaller even terms are also bounded.

Lemma 6 *If $G_{se} > \ln(10)(k-2) + \ln(4)$, $G_{mc} > (2G_{se} - \frac{1}{2k-3})$ and l is even, then $N_{l+2} < \frac{1}{2} N_l$.*

Proof The proof for this sub-lemma is similar to that for Lemma 5, except that the condensation function is defined so that the presence of a double tile in the $l+1^{\text{st}}$ and $l+2^{\text{nd}}$ columns is taken into account.

Here, we use a procedure that takes an assembly A of width $k-1$ and length $l+2$, and then condenses its right end to yield an assembly B with width $k-1$ and length l . Again, A and B are identical except for the rightmost three columns of A and the last column of B ; there, if A had a tile in any of the last three columns in some particular row, then B will have a tile in its last column in the same row. An added detail is that we must now consider that the rightmost two columns of A may contain a double tile; in this case, the rightmost two columns of B must have a double tile also.

The double tile may either be on the top or on the bottom; without loss of generality, we assume it is on the bottom, since the other case can be treated identically. Again, the tile types of the new tiles in B are determined by their neighbors. Formally, we say that $B = \mathbf{condensation}'(A)$ if

$$\begin{aligned} \forall 0 \leq a < k-1, 0 \leq b < l-1 \text{ and } (a,b) \neq (k-2, l-2) : \tilde{A}(a,b) &= \tilde{B}(a,b), \text{ and} \\ \forall 0 \leq a < k-1 : \tilde{B}(a, l-1) = 0 \text{ iff } \tilde{A}(a, l-1) &= \tilde{A}(a, l) = \tilde{A}(a, l+1) = 0, \text{ and} \\ \tilde{B}(k-2, l-2) = 0 \text{ iff } \tilde{A}(k-2, l-2) &= \tilde{A}(k-2, l) = 0. \end{aligned}$$

The proof that every assembly A has a connected $\mathbf{condensation}'$ is virtually identical to the proof in the previous lemma. The rest of the proof is also similar, except that different numbers of tiles may be removed from the $(l+1)^{\text{st}}$ and $(l+2)^{\text{nd}}$ columns.

For a given assembly A , creating \tilde{A} from \tilde{B} , where $\mathbf{condensation}'(A) = B$, requires adding p tiles, $1 \leq p \leq 2k-3$, to the $(l+1)^{\text{st}}$ and $(l+2)^{\text{nd}}$ columns of B , and then removing h tiles, $1 \leq h < k-1$, from the l^{th} column.

For each of the $p-h$ tiles that are added to the $(l+1)^{\text{st}}$ column and $(l+2)^{\text{nd}}$ columns where the corresponding tiles in the l^{th} column are not removed, $p-h$ tiles added to the assembly and no more than $2(p-h)-1$ bonds may be formed. For the h tiles that are moved from the l^{th} to the $(l+1)^{\text{st}}$ column or $(l+2)^{\text{nd}}$, no tiles are added, and no more bonds can be created.

Thus, the spurious nucleation rate of these assemblies is given by:

$$\begin{aligned} N_{l+2} &= \sum_{\substack{A+t \rightarrow C+t \in R_k^{\text{in}} \\ \text{s.t. } \text{length}(A)=l+2}} k_f[A]_{ss} e^{-G_{mc}} \\ &< \sum_{\substack{A,B \\ \text{s.t. } \mathbf{condensation}'(A)=B \\ \text{length}(B)=l}} l_A k_f[A]_{ss} e^{-G_{mc}} \\ &< \sum_{\substack{B \text{ s.t.} \\ \text{length}(B)=l}} \sum_{p=1}^{2k-3} \binom{2k-3}{p} \sum_{h=0}^{k-2} \binom{k-2}{h} l_A k_f[B]_{ss} e^{-G_{mc}} e^{-(p-h)G_{mc}} e^{(2(p-h)-1)G_{se}}. \end{aligned}$$

When $G_{mc} > 2G_{se} - \frac{1}{2k-3}$, this similarly reduces to

$$\begin{aligned} &< \sum_{\substack{B \text{ s.t.} \\ \text{length}(B)=l}} l_A k_f[B]_{ss} e^{-G_{mc}} e^{-G_{se}} (1 + e^{\frac{-1}{2k-3}})^{k-2} (1 + e^{\frac{1}{2k-3}})^{2k-3} \\ &< \sum_{\substack{B \text{ s.t.} \\ \text{length}(B)=l}} l_A k_f[B]_{ss} e^{-G_{mc}} e^{-G_{se}} \left((1 + e^{\frac{-1}{2k-3}})(1 + e^{\frac{1}{2k-3}})^2 \right)^{k-2}. \end{aligned}$$

For $k > 2$, $(1 + e^{\frac{-1}{2k-3}})(1 + e^{\frac{1}{2k-3}})^2 < 10$, and thus

$$< \sum_{\substack{B \text{ s.t.} \\ \text{length}(B)=l}} l_A k_f[B]_{ss} e^{-G_{mc}} e^{-G_{se}} 10^{k-2}.$$



Figure 3.13: **Hypothesized critical nucleus for most spurious nucleation reactions.** The rate of spurious nucleation reactions by this assembly (shown in different shades of gray for tile sets of widths 3,4,5,and 6) accounts for a large portion of spurious nucleation at slow speeds, and also accounts for the rate of increase in spurious nucleation rates as assembly gets faster.

Therefore, when $G_{se} > \ln(10)(k - 2) + \ln(4)$, and recalling that $l_A \leq l_B + 1$,

$$\begin{aligned}
 &< \sum_{\substack{B \text{ s.t.} \\ \text{length}(B)=l}} \frac{1}{4}(l_B + 1)k_f[B]_{ss}e^{-G_{mc}} \\
 &< \sum_{\substack{B \text{ s.t.} \\ \text{length}(B)=l}} \frac{1}{2}l_B k_f[B]_{ss}e^{-G_{mc}} \\
 &= \frac{1}{2} \sum_{\substack{A+t \rightarrow B+t \in R_k^{in} \\ \text{s.t. length}(A)=l}} k_f[A]_{ss}e^{-G_{mc}} = \frac{1}{2}N_l.
 \end{aligned}$$

Now, we can combine Lemma 5 and Lemma 6 to derive Lemma 4. If l is even,

$$\begin{aligned}
 \sum_{p=l+1}^{\infty} N_p &= N_{l+1} + N_{l+2} + N_{l+3} + N_{l+4} + \dots \\
 &< \frac{1}{2}N_l + \frac{1}{2}N_l + \frac{1}{4}N_l + \frac{1}{4}N_l + \dots \\
 &< 2N_l.
 \end{aligned}$$

Thus, to calculate the spurious nucleation rate up an accuracy of $\frac{1}{\delta}$, it is only necessary to compute the inner sums of the series until the sum the current value of l is (even and) less than $\frac{1}{2\delta}$. (Note that this approach does not directly yield a proof of an asymptotic bound for arbitrary k , because the formula for the nucleation rate is not a closed form expression.)

We have used this series truncation method to calculate the rate of spurious nucleation to 2 parts in 10^5 for $k = 3, 4, 5, 6$ and for a range of G_{se}, G_{mc} for which $\epsilon = 0.1$. The values of G_{se}, G_{mc} , and k used were in a regime in which Lemma 4 applies. The results are shown in Figure 3.11.

In addition to the numerical calculations providing lower estimates, the slopes of $\log n_k^+$ vs. $\log r_k$ in Fig. 3.11 are larger than those of $\log n_k^1$ vs. $\log r_k$. Specifically the numerical calculations give slopes $\frac{k+2}{2}$, compared to the asymptotic bounds that give slopes $\frac{k}{2}$. Is this reasonable? In the limit as $G_{mc} \rightarrow \infty$, all spurious nucleation should be dominated by the single species with the highest steady-state concentration (adding tiles become so unfavorable that other species can be neglected).

Tile set width	Steady state	Typical reaction	50 tile assemblies in a typical reaction
3	10^{11} years	$< 9 \times 10^{10}$ years	< 5000 years
4	40 years	< 40 years	< 30 days
5	10 days	< 20 days	< 10 hours
6	7 hours	< 20 hours	< 2 hours

Table 3.1: **Time needed to grow 1000 layers such that less than 1 percent of assemblies are spuriously nucleated.**

The analysis in Section 3.4 suggests that this assembly is the one shown in Figure 3.13. The steady-state concentration of this assembly A for a tile set of width k is $[A]_{ss} = e^{-(2k-3)G_{mc}+(3k-6)G_{se}} = e^{-kG_{se}+(2k-3)\epsilon}$. If all forward nucleation reactions involve A , then

$$n_k^+ = 2k_f[A]_{ss}e^{-G_{mc}} = 2k_f e^{-(k+2)G_{se}+(2k-4)\epsilon}$$

while the speed of growth is

$$r_k = k_f e^{-2G_{se}} \frac{e^\epsilon - 1}{k - 1},$$

and thus the slope would be $\frac{k+2}{2}$, as observed. Even for the range of smaller G_{mc} for which numerical calculations were performed, this estimate of n_k^+ is based on assuming a single critical species is within a factor of three of the precisely calculated value.

3.6.2 Stochastic Simulations for Estimating Nucleation Rates Before Steady State is Achieved

In order to determine whether steady state is a good approximation for what happens in a typical spurious nucleation reaction, we simulated zig-zag tile assembly for tile sets of widths $k = 3, 4, 5$, and 6 and measured the rates of spurious nucleation events during the time it should take to grow 1000 layers from seeds. Since there are an infinite number of powered accretion reactions, exact simulation of growth under the kTAM using mass action dynamics is not possible. Instead, we simulated assembly growth using stochastic chemical reaction dynamics. To approximate the nucleation rate, we simulate a tiny reaction volume, and use these results to predict the nucleation rate in a much larger volume.

We used the Gillespie algorithm [Gil76] to sample the trajectories of stochastic dynamics of the zig-zag tiles in a small volume V , chosen to ensure accuracy as described below. Following the powered model, our simulation assumes the concentration of each tile type to be constant and explicitly tracks each assembly containing more than one tile. Initially, no multi-tile assemblies are present. Single tiles are present at a concentration of $e^{-G_{mc}}$ so that the rate of two tiles colliding (and thus producing a new assembly to be explicitly tracked) is $\mathbf{A}k_f V e^{-2G_{mc}}$ molecules / second, where \mathbf{A} is Avogadro's number. For each assembly containing two or more tiles, the rate of tile addition is $k_f e^{-G_{mc}}$ and the rate that a tile with b bonds falls off an assembly is $k_f e^{-bG_{se}}$.

For $k = 3, 4, 5, 6$ and a range of G_{se} and G_{mc} where $\epsilon = 0.1$, we counted the number of spurious nucleation events, m , that took place over the time course of a “typical reaction”, $s = 1000/r_k$, in a volume V that was chosen large enough to ensure that statistical error in m is less than 10 percent of its value ($P > 0.95$). If our simulations yield a nucleation rate of m events per second, the molar rate of nucleation events for a bulk volume is given by $n_k^+ \approx \frac{m}{V\mathbf{A}}$. The results of the simulation—which were possible only for small enough G_{se} such that nucleation events were frequent enough to be counted—are shown in Figure 3.12. For $k = 3$ and $k = 4$, these rates are within a factor of 2, and all values tested are within a factor of 10 of the linear extrapolation of the curves from Fig. 3.11, indicating that the choice in Section 3.5 to bound nucleation rates based on steady-state concentrations did not affect our estimate of nucleation rates too greatly. This should be expected, given that under the conditions we studied most steady-state nucleation appears to involve assemblies like the one shown in Figure 3.13.

3.6.3 Nucleation of Long Ribbons

In this paper, we have defined a spurious nucleation reaction for a zig-zag tile set of width k as a reaction in which an assembly of width $k - 1$ grows to width k . The goal was that this definition would be inclusive, such that all long ribbons would undergo at least one spurious nucleation reaction, but not too loose, such that most spurious nucleation reactions lead to a long ribbon. However, many of these spurious nucleation reactions are not energetically favorable—an assembly may briefly reach width k before a tile falls off. The assembly then either melts or undergoes another spurious nucleation reaction.

At what rate do long ribbons appear? Using the stochastic simulation described in the last section and the same range of physical reaction parameters, we measured m' , the number of ribbons containing 50 tiles or more that were present at the end of a “typical reaction”, for the widths 3, 4, 5, or 6. This counts the number of spurious nucleation events that did *not* subsequently melt, and thus it provides the basis for an estimate for n_k . As only those crystals that nucleated sufficiently far before the end of the simulation will have grown to a large enough size to have been counted, so we use the formula $n_k \approx \frac{m'}{(s-50/(k-1)/r_k)V\mathbf{A}}$. The results are shown in Figure 3.12.

While the assumption that concentrations of unnucleated assemblies had reached steady state appears to have been inconsequential, these simulations indicate that the assumption that a spurious nucleation reaction always produces a long ribbon did cause a significant overestimate of the amount of spurious nucleation. They suggest that most (at least 99%) of assemblies that reach critical size will subsequently melt.

3.6.4 Expected Effectiveness in Practice

Do these results indicate that nucleation control with tile sets of width 6 or less are good enough? Recall that our “reasonable goal for a typical reaction” addresses how much time is needed to grow seeded ribbons of 1000 layers with less than 1% of the crystals being spuriously nucleated. The fraction of crystals that are spuriously nucleated is given by $f = \frac{L}{c} \frac{n_k}{r_k}$, where L is the number of layers to be grown on seeds, and c is the concentration of seeds. While the simulations only measured n_k for large values of r_k , it is possible to bound n_k from above for smaller values of r_k by linearly extrapolating the lines in Figure 3.12, because the slopes $\log n_k$ vs $\log r_k$ should be no smaller than at steady state, $\frac{k+2}{2}$. Taking $c = \frac{1}{L} e^{-G_{mc}}$, we use this technique to bound r_k from above, and therefore to bound the time necessary to grow 1000 layers on average where less than

1% of the crystals are spuriously nucleated. The results, shown in Table 3.1, are encouraging. They suggest that, using a zig-zag tile set of width 6, just a few hours would be enough to avoid most spurious nucleation.

The analysis and simulations in this section support the idea that nucleation control using the zig-zag tile set not only works, but is practical. While in most respects our models appear complete, two effects which may be important in the actual process of assembly are not included. One important effect is tile depletion: while our model considers the concentration of free tiles to be constant, in a typical experiment tiles are used up because they join assemblies. Since the rate of spurious nucleation is concentration dependent, we would expect the rate of spurious nucleation to be larger at the beginning of a reaction, when almost all free tiles remain, than at the end, when many tiles are used up. Because of this effect, our simulations may actually *overestimate* the spurious nucleation rate.

However, our simulations also neglect an important possible reaction pathway that may greatly increase the rate of spurious nucleation. While our model assumes tiles must be added to assemblies one at a time, in an experiment, small assemblies can also attach to each other. The formation and joining of several small assemblies may be faster than the spurious nucleation pathways described in this paper. A complete understanding of spurious nucleation of zig-zag tiles requires an understanding of the speed of spurious nucleation reactions caused by aggregation.

3.7 Conclusions

3.7.1 Nucleation of Algorithmic Self-Assembly

Our original motivation for this work was to show that self-assembly programs that work in the aTAM, in which it is straightforward to design tile sets that algorithmically assemble any computationally defined structure, can also be made to work in the more realistic kTAM. Tiles sets that assemble correctly via unseeded growth in the aTAM with a threshold of $\tau = 1$ will assemble correctly in the kTAM under the right conditions. However, tile sets that are designed to assemble via seeded growth in the aTAM with a threshold $\tau = 2$ may fail in the kTAM because mismatch, facet, and spurious nucleation errors occur. These problems are ameliorated in the limit of slow assembly speed [Win98]. Other work has described methods to control mismatch errors and facet errors without significant slowdown [WB04, CG05, RSY05]. Here, we have developed a construction that corrects the last discrepancy, spurious nucleation errors, again without significant slowdown.

However, it remains to be formally proven that these constructions can be combined to control all types of errors simultaneously for any tile set of interest. No major difficulties are expected, in large part because mismatch and facet errors can both be controlled by a single mechanism [CG05] and the control of spurious nucleation errors works independently of this mechanism. Both methods work by transforming an original tile set which works in the aTAM at $\tau = 2$ into a new (typically larger) tile set that is more robust to particular kinds of errors in the kTAM. Additionally, the combination of these error correction mechanisms is expected to be experimentally tractable: the cost of both these transformations is a moderate increase in spatial scale and the number of tile types.

We expect that the zig-zag tiles can be used as a subroutine in more complex algorithmic self-assembly programs when control of nucleation is needed. Other self-assembly programs for controlling nucleation are certainly possible; we do not know whether the zig-zag tile sets are optimal.

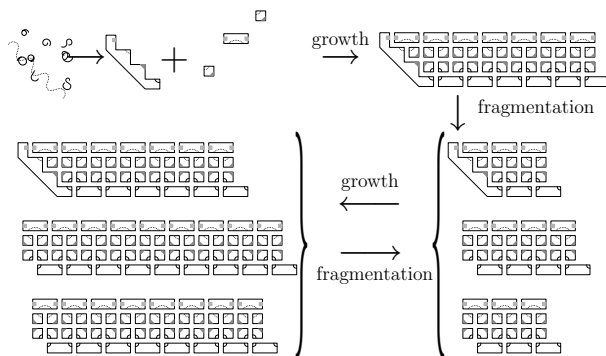


Figure 3.14: **Exponential amplification of assemblies.** Probe strands assemble onto a target sequence to create a seed assembly, which nucleates zig-zag growth. Periodic fluid shear causes fragmentation of zig-zag assemblies, leading to exponential amplification. The diagonal structure of the seed assembly shown here is a natural shape for assembling tiles on a scaffold strand [RPW04].

3.7.2 Detection of a Single DNA Molecule

Control over nucleation in algorithmic self-assembly can be seen as a special case the detection of a single molecule. For a tile set of sufficiently large width, essentially nothing happens when no seed tiles are present, whereas if even a *single* seed tile is added, growth by self-assembly will result in a macroscopic assembly. Theorem 1 shows that the *false-positive* rate for detection can be made arbitrarily small by design; the *false-negative* rate in the kTAM is approximately 0. Although this idealized model does not consider many factors that could lead to poorer detection in a real system, we don't know of any insurmountable problems with implementing single-molecule detection this way.

There are, however, two immediate drawbacks. First, detecting seed-tile assemblies is not as useful as detecting arbitrary DNA sequences. Second, the linear growth of a single zig-zag assembly would require a long time lapse before a macroscopic change is perceptible. As sketched in Figure 3.14, we can surmount both obstacles. First, as in [MLRS00a, YLFR03], a set of strands can be designed to assemble double-crossover molecules on a (sufficiently long) target strand with nearly arbitrary sequence, thus creating the seed assembly if and only if the target strand exists. Second, since fluid shear forces can fragment large DNA assemblies, intermittent pipetting or vortexing could break large zig-zag assemblies, increasing the number of growing ends with each fragmentation episode. This fragmentation process can be expected to lead to exponential growth in the number of zig-zag assemblies without increasing the false-positive rate. (When a spuriously nucleated assembly does eventually form, of course, it will also be exponentially amplified.)

Based on the analyses of the previous section, we can estimate the effectiveness of this procedure. Is there a reasonable tile set width for which a single seed could amplify to a level of detectability in a reasonably short time without any spurious nucleation occurring within the given volume? Specifically, given a $10 \mu\text{L}$ reaction volume, a minimum detection level of 10^5 crystals and a protocol in which assemblies split after growing on average to size 200 layers, we would like to determine the minimum time and tile set width that meet these requirements. Creating 10^5 crystals requires first growing from the seed to size 200, then 17 cycles of fragmentation followed by growing 100 additional layers (50 on each side), so amplification requires $t_a = 850/r_k$ seconds. The expected time for the first nucleation event is $t_f = \frac{1}{n_k V \mathcal{A}}$, and our criteria for reliable detection is $t_f > 100t_a$,

i.e., $\frac{n_k}{r_k} < \frac{1}{85000V\mathcal{A}}$. Based on Figure 3.12, we use the approximation $\log(n_k)+2 = \frac{k+2}{2}(\log(r_k)-1.7)$. Solving for t_a as a function of k , we find that good results are obtained for experimentally feasible widths. For example, with $k = 12$, reliable detection of a single seed in $V = 10 \mu\text{L}$ is $t_a \approx 26$ hours.

Chapter 4

Preventing Spontaneous Nucleation in the Laboratory

Abstract

A central goal of chemistry is to fabricate supramolecular structures of defined function and composition. In biology, control of synthesis is often achieved through precise control over nucleation and growth processes: a seed molecule initiates growth of a structure, but in the absence of a seed, growth is inhibited by a large kinetic barrier. Here, we show how such control can be systematically designed into self-assembling supramolecular structures made of DNA tiles, called zig-zag ribbons. Ribbons have a fixed width, chosen by design, but can elongate indefinitely. Theory predicts that under slightly supersaturated conditions, elongation is favorable but the energetic barrier to nucleation of a new ribbon is proportional to its width. Here, we show experimentally that while zig-zag ribbons of different widths have similar thermodynamics, nucleation rates decrease for wider ribbons, indicating that we can program the nucleation rate by choosing a ribbon width. The presence of a seed molecule, a stabilized version of the presumed critical nucleus, removes the kinetic barrier to nucleation of a ribbon. Thus, we demonstrate the ability to grow supramolecular structures from rationally designed seeds, while suppressing spurious nucleation. Control over DNA tile nucleation, by programming a kinetic pathway in algorithmic crystal growth, will make possible the high yield synthesis of micron-scale structures with complex programmed features. More generally, this work shows how a subroutine for self-assembly can be initiated.

4.1 Introduction

Biology demonstrates the potential of self-assembly to create sophisticated organization on the molecular scale. The fundamental challenge in engineering novel self-assembled objects is that the molecular components must themselves contain the information needed to guide self-assembly.

Controlling the nucleation of a self-assembled object is the first step to controlling the self-assembly process, as exemplified by the formation of actin networks [WM02, SM01], the growth of microtubules on the centrosome [MBS⁺95], and the assembly of bacterial flagella [AH02]. These systems avoid the difficulty of controlling homogeneous nucleation by relying on heterogeneous nucleation: growth is rare except in the presence of a seed molecule, from which it proceeds with little or no kinetic barrier. Here, we demonstrate a general design strategy for creating self-assembled

molecular structures for which nucleation is similarly controlled by a seed. We use programmable DNA tiles to create a series of seeded “zig-zag ribbons” that exhibit increasing kinetic barriers to homogeneous nucleation.

DNA tiles [FS93, LYQS96] are a general-purpose nanoscale construction material. A DNA tile consists of multiple strands woven together to form double helices connected by “crossover-points” (Figure 4.1a). Tiles interact through the hybridization of their single-stranded (“sticky”) ends (Figure 4.1b) and can assemble into extended structures, including 1- and 2-dimensional crystals [WLWS98, LYK⁺00, RENP⁺04, MLK⁺05]. The interactions between tiles are programmed through the design of sticky ends; complementary sticky ends hybridize, while non-complementary sticky ends are unlikely to interact. Under slightly supersaturated conditions, where the attachment of a tile to an assembled crystal by two or more sticky ends is favorable but attachment by just one sticky end is unfavorable, in principle it is possible to program complex assembly processes [Win96].

We have constructed sets of DNA tiles that assemble to form ribbons of particular widths. Ribbon assembly proceeds in two phases: nucleation and growth. Under slightly supersaturated conditions, nucleation requires a mixture of favorable and unfavorable tile attachments. In contrast, growth proceeds through a series of favorable monomer tile addition reactions (Figure 4.1d). The crystals are designed such that the number of unfavorable attachments required for nucleation is exactly the width of the ribbon minus one. Theoretically, increasing the number of required unfavorable reactions can exponentially reduce the rate of nucleation [SW05a].

In this paper, we describe the design and synthesis of 3, 4, 5, and 6 tile-wide ribbons, denoted ZZ3 – ZZ6. We show that each type of ribbon forms as designed and that there is a kinetic barrier to homogeneous nucleation of ribbons of all widths. The growth rates of ribbons under slightly supersaturated conditions confirm that nucleation rates are lower for wider ribbons. Finally, we synthesize a seed molecule for ZZ4 ribbons and show that the kinetic barrier to nucleation is greatly reduced in its presence.

4.2 Materials and Methods

Design. DAO-E tiles (Figure 4.1a,b) were used to construct ribbons [WLWS98]. Sequences for the double-stranded tile regions were either those used previously [RPW04] or were designed by computer using sequence symmetry minimization [See90, DLWP04], a technique that selects sequences with minimum undesired intramolecular and intermolecular interactions. Sticky end sequences were designed to minimize differences in binding strength between pairs of complementary ends and to minimize the binding energy between non-complementary ends. Binding energies were estimated by the nearest neighbor energy model [San98]. Details are given in the Supplementary Information.

Experiments. All reactions were performed in Tris-Acetate EDTA buffer to which 12.5 mM hydrous MgCl₂ was added. Non-denaturing gel electrophoresis showed that all tiles formed single products with at least 80% yield and that crystal seeds formed with roughly 50% yield. UV absorbance was measured in an AVIV 14DS spectrophotometer (AVIV Biomedical, Lakewood, NJ) equipped with a computer-controlled temperature bath. Atomic force microscopy (AFM) was performed on a Digital Instruments Nanoscope III (Veeco Metrology) in fluid tapping mode using NP-S tips. See the Supplementary Information for details.

Modeling. The simplified model of zig-zag ribbon nucleation shown in Figure 4.1d was expressed as a set of chemical reactions. This “the standard sequence” model (for nucleation and growth) considers A_n an assembly containing n single or double tiles. The model ignores the com-

binatorial number of possible species of each size. A_n has a standard free energy $\Delta G_n^\circ = b_n \Delta G_{se}^\circ$ where b_n is the number of sticky end pairs formed in assembly A_n (e.g., $b_6 = 7$ for the top assembly in Figure 4.1d) and $\Delta G_{se}^\circ = \frac{1}{2}(\Delta H^\circ - T\Delta S^\circ)$. Values for ΔH° and ΔS° refer to the enthalpy and entropy of a tile attaching to a ribbon by two sticky ends, as determined by the experiments described below. The free energy of tile attachment by a single sticky end is unknown; $\frac{1}{2}(\Delta H^\circ - T\Delta S^\circ)$ is used as a simplifying assumption. The model includes all reactions of the form $A_n + A_m \rightleftharpoons A_{n+m}$ for $1 \leq n \leq m$ such that $n+m \leq M$, where M is the largest ribbon size modeled and is limited by computation time to 100. These reactions include growth by monomer addition as well as joining and internal scission of ribbons [ENKF04]. As a simplification of size-dependent diffusion-limited reaction rates, we use a forward constant $k_f^{n,m} = k_f = 10^6$ /M/s (typical for oligonucleotide hybridization [Wet91]) for reactions involving a tile or pre-critical nucleus, while for reactions involving two ribbons $k_f^{n,m} = k_j = 15000$ /M/s as estimated experimentally (see below). The reverse rates are determined by thermodynamics: $k_r^{n,m} = k_f^{n,m} e^{-(\Delta G_n^\circ + \Delta G_m^\circ - \Delta G_{n+m}^\circ)/RT}$. There are no fitting parameters.

4.3 Results

Ribbon Structure. To verify that the tiles assembled into the designed structures, the strands for each ribbon were annealed at 100 nM (per strand) from 90 °C to 20 °C over 20 hours in a PCR machine. Atomic force microscopy of each sample showed predominantly the desired structures (Figure 4.2). Most ribbons were microns (hundreds of tile layers) long, suggesting that ribbon nucleation was much slower than growth. To establish that annealing is necessary to produce long ribbons, we mixed pre-formed tiles at room temperature, where nucleation is presumed to be fast, and let them sit for 20 hours. Few ribbons longer than 10 tile layers were observed.

Homogeneous nucleation rates. The rate and temperature dependence of zig-zag ribbon crystal growth and melting was studied using ultraviolet spectroscopy. At 260 nm, single-stranded DNA absorbs more light than double-stranded DNA, so changes in absorbance provide a quantitative measure of hybridization. We annealed and melted samples of the strands for zig-zag ribbons from 90 °C to 20 °C and back to 90 °C at approximately 0.13 °C per minute in a spectrophotometer (AVIV 14-DS, Lakewood, NJ) to determine formation and melting temperatures of the ribbons. There was a reversible absorbance change between 90 °C and 45 °C and a region of hysteresis between 40 °C and 25 °C (Figure 4.3a). The previously determined melting temperatures of DNA tiles and DNA tile assemblies [Rot01, RPW04] suggested that the reversible higher-temperature transition was due to formation and melting of tile cores from/into strands, while the hysteretic transition was due to formation and melting of zig-zag ribbons from/into tiles.

To confirm that the hysteretic transition was the result of ribbon formation and melting, we also studied a set of tiles lacking sticky ends but otherwise identical to the original ribbon tiles. Anneals and melts of ribbon tiles with and without sticky ends had the same shape in the region between 90 °C and 50 °C, but no significant absorbance changes at lower temperatures were observed for the tiles lacking sticky ends (Figure 4.3a). Thus, the absorbance changes in the regime between 40 °C and 25 °C mostly reflect ribbon formation.

We therefore tabulated the melting and formation temperatures, defined as the temperatures where half the ribbons were melted or formed, for 4 concentrations of each of the 4 ribbon widths (Figure 4.3b–d). For a given ribbon type and concentration, the amount of hysteresis is dependent on the speed of the melt: at equilibrium, the formation and melting temperatures are the same. As

the speed of annealing and melting increase, conditions diverge from equilibrium, and the differences between these temperatures grow larger. At the speed at which we performed the melts, the melting temperatures of all the mixtures were approximately the same, but the formation temperatures were strongly concentration dependent and very slightly dependent on ribbon width.

While ZZ3–ZZ6 were designed so that the melting temperatures of ZZ3–ZZ6 at a single concentration were the same, the predicted energies of sticky ends interactions suggested that the melting temperature of ribbons at the highest and lowest concentrations should vary by several degrees. This prediction is valid in a regime where melting occurs by attrition of one tile at a time from the end of the ribbons (Figure 4.1d). Because the ribbons are long, such attrition is slow. Work on DNA nanotubes found that the melting of the nanotubes was greatly sped up by tube scission during melting [ENKF04], which suggested that the melting in the Figure 4.3 experiments may also have occurred primarily by scission.

Ribbons likewise grow by adding one tile at a time. (In contrast, growth of 2- and 3-dimensional crystals is faster, and growth speed increases as the crystal grows.) Thus, the speed the sample was annealed at may also have been fast relative to growth and nucleation rates. The measured formation temperature was likely the temperature at which the kinetic barrier to nucleation largely disappears and many small nuclei form. These small nuclei could then join to form long ribbons. Under these conditions, the formation temperature would be dependent on the concentration of the tiles, but not the width of the ribbon, as we observed.

We therefore designed a second set of experiments with the goal of measuring ribbon nucleation and growth rates under conditions where there is a kinetic barrier to nucleation. Instead of annealing and melting the samples at a continuous rate, these experiments hold the sample at one temperature for many hours. Samples were cooled from 90 °C as before, to a target temperature between 25 °C and 41 °C. When the target temperature was reached, the sample was held at this temperature for 24 hours, then cooled to 15 °C, where ribbons are fully formed. The samples were then melted back to the target temperature at the original speed, then held for another 24 hours. All heating and cooling was done at the same speed as the melts shown in Figure 4.3.

For these “temperature-hold” experiments, we held samples of ZZ3, ZZ4, and ZZ6 at temperatures between 25 °C and 41 °C. Absorbances from three ZZ4 experiments are shown in Figures 4.4a–c. At 39 °C, above the melting temperature, there is no activity apart from an initial transient attributed to tile formation. At 33 °C, there is a large barrier to nucleation: after 24 hours, there is still a significant separation between the anneal and melt. At 25 °C, there is a smaller kinetic barrier to nucleation: the absorbances of the anneal and melt holds converge within 24 hours. The traces of all melt holds show no significant change after 6–12 hours.

To study growth rates during the experiments, the normalized absorbances at the beginning and end of each hold are plotted in Figures 4.4d–i. The blue and cyan lines recapitulate the absorbance values during the temperature ramp. The red line reflects the amount of growth seen during the anneal hold, while the magenta line similarly reflects changes during the melt hold. The red and magenta lines converge above 35–36 °C because no ribbons form at these temperatures (the absorbance changes during the anneal hold are assumed to be due to tile formation). The black dotted lines trace our fit of the absorbance values corresponding to unbound tiles, reflecting the slight temperature dependence of the absorbance of double stranded DNA [MB87].

Hysteresis remaining after 24 hours (as measured by the gray area) grows with increasing ribbon width and with decreasing concentration. Formation and melting temperatures were defined analogously to the temperature-ramp experiments (Figure 4.4d–k). The increased hysteresis is

primarily due to changes in formation temperatures, as the melting temperature for temperature-hold experiments exhibits no measurable dependence on ribbon width, and only a slight dependence on concentration. This data suggests that seeded growth of wider ribbons can proceed without significant spontaneous nucleation over a wide temperature range, and is suggestive of a larger kinetic barrier to nucleation for wider ribbons.

Standard sequence simulations qualitatively reproduced the hysteresis, and formation and melting temperatures as a function of ribbon width and concentration, although in simulation hysteresis was consistently more pronounced. As predicted by the model, ribbons at higher concentrations melted at higher temperatures. Equilibrium tile concentrations were achieved within a few hours of holding during the melt in the simulations. Because absorbance values also quickly reached their maximum values in all melt holds performed, we infer that equilibrium was achieved by the end of the melt holds.

The dependence of T_m on concentration can be used to estimate the energetics of tile attachment to ribbons. At the melting temperature, the free energy of tile attachment, $\Delta G = \Delta G^\circ - nRT \ln([m])$, is zero. Here, $\Delta G^\circ = \Delta H^\circ - T\Delta S^\circ$ is the standard free energy of formation for a tile attaching by two sticky ends. Fitting the data using a van't Hoff plot [DB02] (Figure 4.4k) gives $\Delta H^\circ = -104.4$ kcal/mol and $\Delta S^\circ = -0.300$ kcal/mol/K. (At 37 °C, $\Delta G^\circ = 9.30 \pm 0.20$ kcal/mol.) Since all three ribbons had comparable melting temperatures, we treated all ribbons as having the same energy for tile attachment, and pooled the data for fitting. The measured ΔH° and ΔS° are comparable to values (-99.6 kcal/mol and -0.265 kcal/mol/K) predicted by the nearest neighbor model of DNA hybridization [BCT00].

Nucleation rates can be estimated from temperature-hold experiments in which tile concentration, $[m]$, decreases only marginally. Although classical nucleation theory [DG00] predicts that the nucleation rate n_r is proportional to $[m]^n$, where n is the number of tiles in the critical nucleus, in these cases n_r will be roughly constant. Since joining was observed at room temperature at a rate of $k_j = 15000$ /M/s, we included ribbon joining in our model. Ribbon joining also explains why anneal holds in the hysteretic regime (e.g., Figure 4.4b) stop changing prior to reaching equilibrium: nucleation is reduced due to its dependence on tile concentration, and few ribbon ends remain to deplete tile concentration. We use

$$\frac{d[r]}{dt} = n_r - k_j[r][r] \quad \frac{d[m]}{dt} = \frac{2}{N}(k_r[r] - k_f[m][r])$$

where $[m]$ is the concentration of each type of tile, $[r]$ is the concentration of ribbons ($[r] = 0$ at $t = 0$), k_f and $k_r = k_f e^{\Delta G^\circ/RT}$ are the rate constants for tile attachment and dissociation, and $N = 2w - 2$ is the number of tile types for a ribbon of width w . All these constants have measured or assumed values; n_r was fit to the data.

We determined n_r at the melting temperatures for each concentration, where the supersaturation $\sigma = \ln([m]_0/[m]_{eq}) = \ln(2)$ is mild. With our best estimates for k_f and k_j (given above), the inferred nucleation rates at 200 nM are 3×10^{-7} nM/s, 9×10^{-7} nM/s, and $> 20 \times 10^{-7}$ nM/s for ZZ6, ZZ4, and ZZ3, respectively. Uncertainties from measurement error, estimation of k_f and k_j , and residual absorbance change due to continued tile formation render absolute values for n_r unreliable. However, for every concentration except 25 nM, the inferred nucleation rates decrease monotonically for wider ribbons, and this conclusion is robust to 10-fold changes in k_f and k_j . Although the effect is not as strong as expected, these results support theoretical predictions that nucleation rates should decrease with width [SW05a].

Heterogeneous nucleation of ribbons. To test whether a segment of full-width ribbon works as a seed for ribbon growth, we designed a stable seed for ZZ4 resembling two layers of tiles in the ribbon. The seed structure has sticky ends identical to those on one side of a ZZ4 repeat unit (Figure 4.1c), but its strands are woven so that it cannot easily fall apart into individual tiles (Figure 4.5a). The seeds form with approximately 50% yield (Figure 4.5b) and have a melting temperature of roughly 62 °C at 2 nM, well above the melting temperatures of the ribbons.

To demonstrate seeded growth, two samples of ZZ4 were annealed from 90 °C to 40 °C, at which point seeds were added to one of the samples (4% by concentration). The samples were then cooled from 40 °C to 34 °C and held at 34 °C for 12 hours. After the hold, the samples were cooled to 15 °C and reheated to 34 °C where the temperature held for another 12 hours (Figure 4.5d). By the onset of the hold, the distance between the anneal and melt was smaller in the sample to which seeds were added. We propose that ribbons had by this time already nucleated on the seeds. The difference in the anneal and melt signals completely disappears after a couple of hours for the seeded sample. Growth remains slow in the seedless sample.

4.4 Conclusions

The results here indicate that it is possible to engineer pathways of crystal growth using rational design. More generally, our methods suggest a way to control the onset of a stage of self-assembly. Used recursively, this control can allow the ordered progress of multiple self-assembly reactions that together produce a complex structure.

Such kinetic control within self-assembly can be surprisingly powerful. It is theoretically possible to design a set of DNA tiles to assemble any computable shape or pattern [Win96, SW04]. This algorithmic crystal growth has been demonstrated experimentally [RPW04, BRW05], but thus far spurious nucleation and errors during growth have resulted in poor yields. It may be possible to use the techniques developed here to reliably grow technologically relevant molecular structures with high yield, including crystals of exact rectangular dimension [RW00, ACGH01] and the layout for a nanoscale demultiplexor circuit or memory circuit [CRW04]. Further, the control of nucleation in ribbons could allow single molecule detection, or the replication and evolution of crystal sequences encoded in ribbons [SW05b].

The approach described here for the control over nucleation is potentially applicable to other organic and macromolecular crystals if sufficient control over intermolecular interactions can be achieved by design. Generalization to two- or three-dimensional assemblies is also possible. Control over nucleation of supramolecular assemblies also can be achieved by conformational changes [DP04] or by energy-consuming enzymatic activity [BZTL02]. These methods, along with the work described here, present the engineer of complex self-assembly processes with a rich design space.

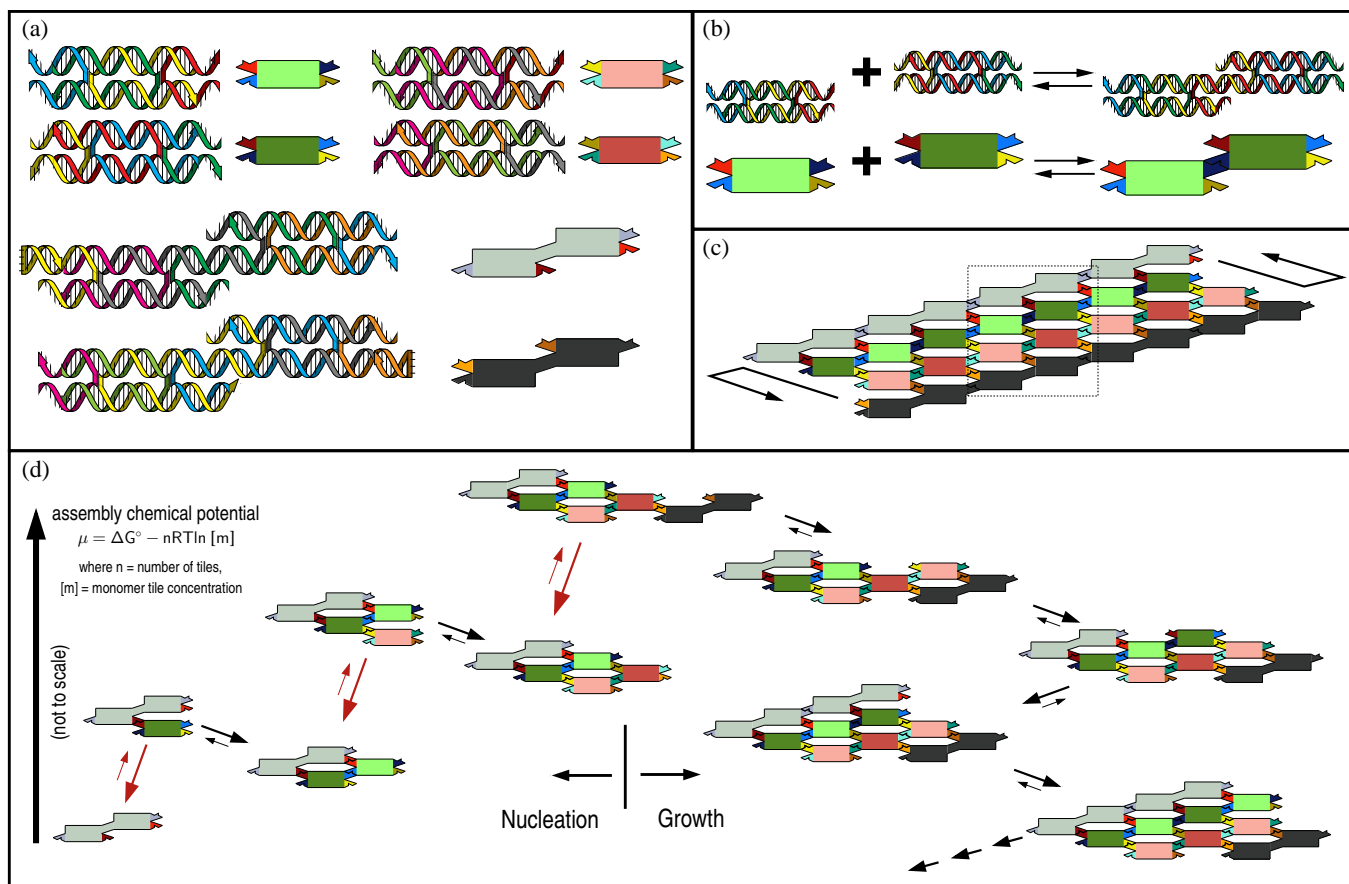


Figure 4.1: Zig-zag ribbon design. (a) Ribbon diagrams of DNA tiles used to construct ZZ4 (4-tile-wide zig-zag ribbon). The 3' end of each strand is indicated by an arrow. A tile consists of 4–6 synthetic DNA strands which form the structures shown because of a preference for hybridization between Watson-Crick complementary subsequences. Tiles have double stranded DNA in the middle (core) of the molecule and single stranded DNA at the ends (called sticky ends). Each tile monomer is either a “single” tile (top four tiles) or a “double” tile (bottom two tiles). A tile can have either 3' or 5' ends on the top as it is oriented within the ribbon structure. Like single tiles, double tiles come in two orientations, but only one orientation is used in ZZ4. The double tiles used here contain inert ends (an uncomplemented sticky end, blunt end, or hairpin) on either the tops or the bottoms of the tiles. Each strand shown has a unique sequence; colors distinguish strands in individual tiles. Single and double tiles of both orientations are depicted by rectangle and claw diagrams shown to the right of the strands. Colors of tile cores distinguish a tile type and claws with the same color represent complementary sticky ends. (b) Tiles bind by hybridization of their sticky ends. Non-complementary sticky-ends tend not to hybridize. (c) The tile structure of ZZ4. A dashed box encloses the 6 tile types in each repeating unit. In principle, growth of a zig-zag ribbon under slightly supersaturated conditions consists mainly of attachments of monomer tiles to ribbons by two sticky ends. The ribbons are designed so that these additions proceed in a zig-zag growth pattern on each side of the ribbon, as denoted by the arrows. (d) One designed pathway for nucleation and growth of ZZ4, consisting of nucleation steps (left) culminating in the critical nucleus (top) followed by growth (right). A monomer tile is added to the growing crystal at each reaction step.

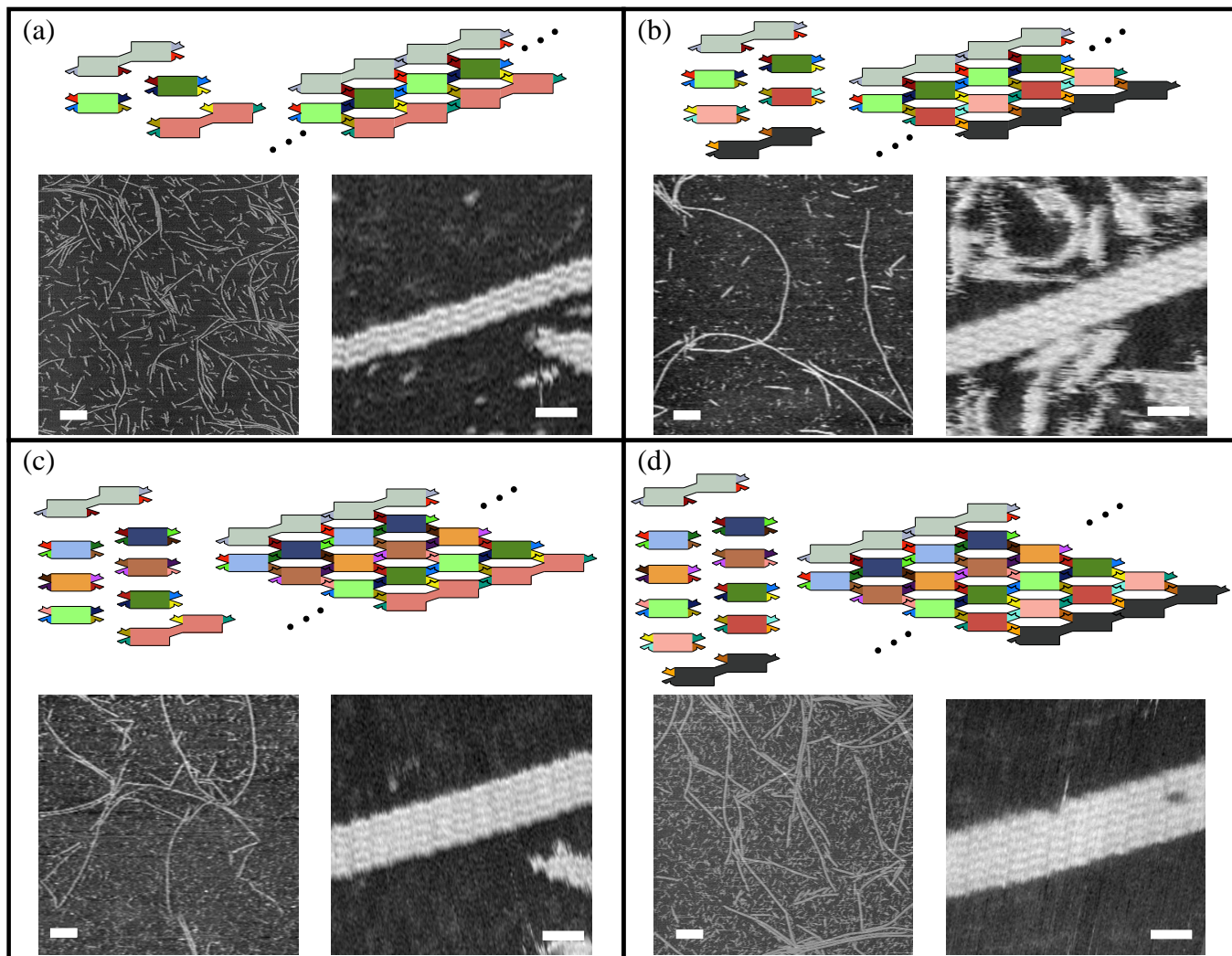


Figure 4.2: **Atomic force microscopy of zig-zag crystals.** Tile sets for ZZ3–ZZ6 and atomic force microscopy images of ribbons (left of (a)–(d)) and their fine structure (right of (a)–(d)). Ribbons sometimes rip during sample preparation, leaving ribbon fragments stuck to the surface. Scale bars are 500 nm (left) and 25 nm (right).

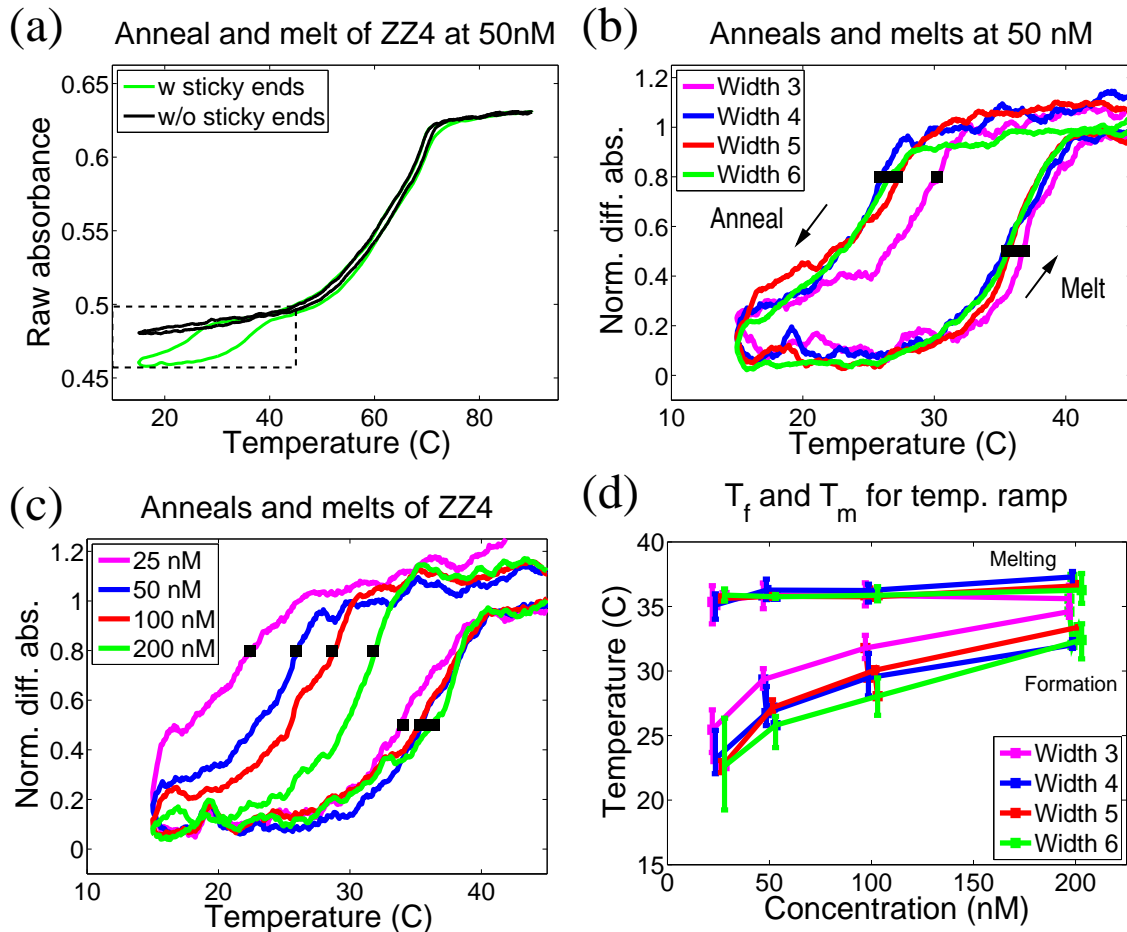


Figure 4.3: **Temperature-ramp anneals and melts of zig-zag crystals.** (a) An example of a UV melt of zig-zag tiles as designed (green) and with sticky ends omitted (black). Because of cuvette variations and changes in stoichiometry that result from mixing so many strands together, the results from the second melt were normalized so that the signals at the 90 °C and at 42 °C are the same. The dashed box encloses the area of the melts shown in (b) and (c). (b) UV melts of ZZ3–ZZ6 at 50 nM in the temperatures where zig-zag crystal formation was observed. Signals shown are the difference between normalized absorbance signals with and without sticky ends. In each loop, the upper lines are annealing curves and the lower lines are melting curves. The formation and melting temperatures, where half the material is formed or melting, respectively, are shown by black squares. (c) UV melts for a single width tile set at different concentrations (25, 50, 100, and 200 nM of each tile). Black squares are as in (b). (d) Formation and melting temperatures (determined as in (b)) for the four widths and concentrations of zig-zag crystals at the four concentrations used in (c). Points at each concentration are staggered so error bars are visible; measurements are for 25, 50, 100, and 200 nM tile concentrations for each ribbon width. Melting temperatures are consistently higher than formation temperatures, indicating hysteresis.

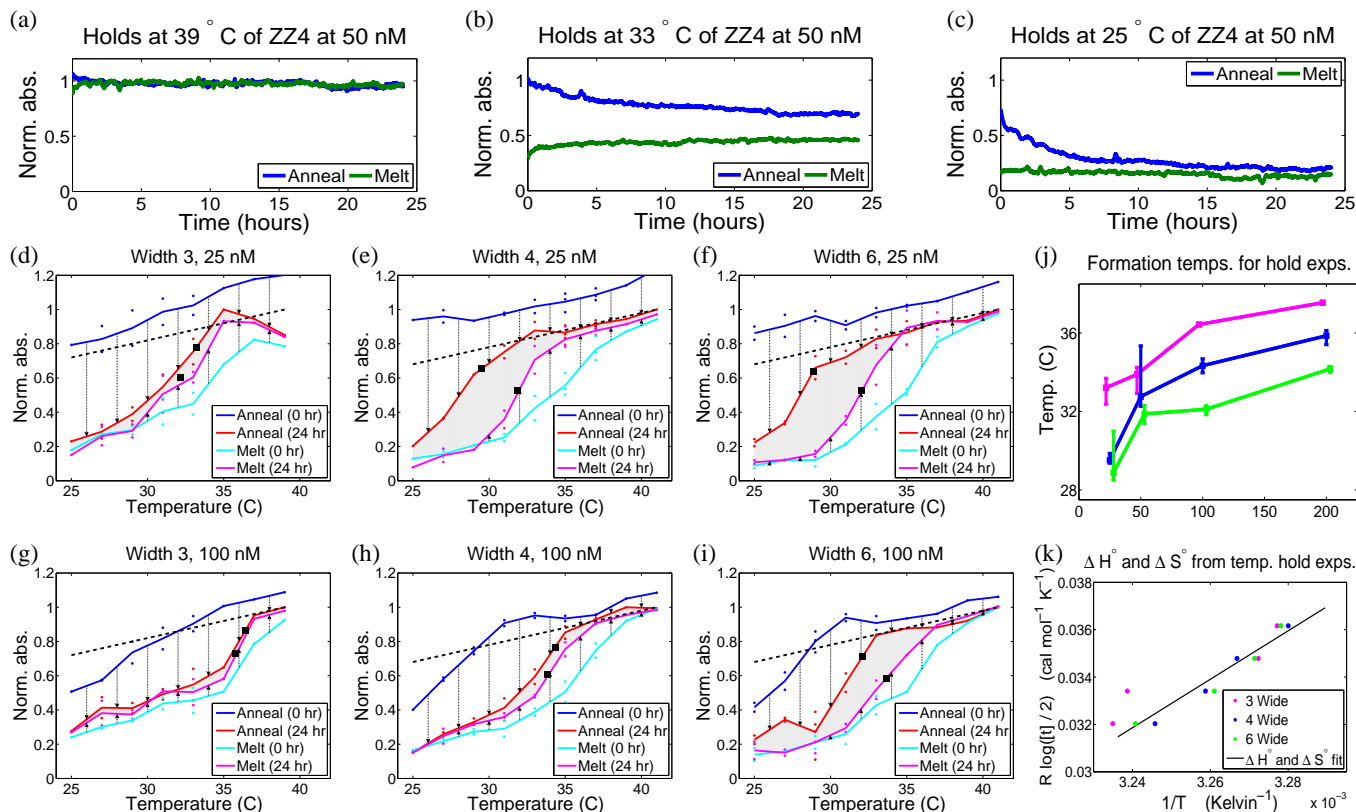


Figure 4.4: **Zig-zag crystal growth and melting at constant temperatures.** (a)–(c) Trajectories showing growth and melting of 50 nM ZZ4 at three different temperatures. Absorbance values are normalized as described in (d)–(i). (d)–(i) Absorbance at the beginning and end of temperature holds. The gray region shows the difference in absorbance after 24 hour holds. Black squares indicate the formation and melting temperatures. Absorbances were normalized so that 1 is the largest absorbance value along the red line, and 0 is the absorbance measured at 15 °C. (j) Formation temperatures determined from temperature-hold experiments. (k) Determination of ΔH° and ΔS° from melting temperatures in temperature-hold experiments, which are presumed to be at equilibrium. The van't Hoff plot shows $1/T_m$ vs. $R \ln([m]_0/2)$, where $[m]_0$ is the initial tile concentration. Because at the melting temperature, the equilibrium free tile concentration is half the initial, and $\frac{1}{2}[m] = K_{eq} = \exp(\Delta H^\circ - T\Delta S^\circ)/RT$, the slope and intercept of this plot determine ΔH° and ΔS° respectively. The outlier (ZZ3, 100 nM) was discarded during fitting.

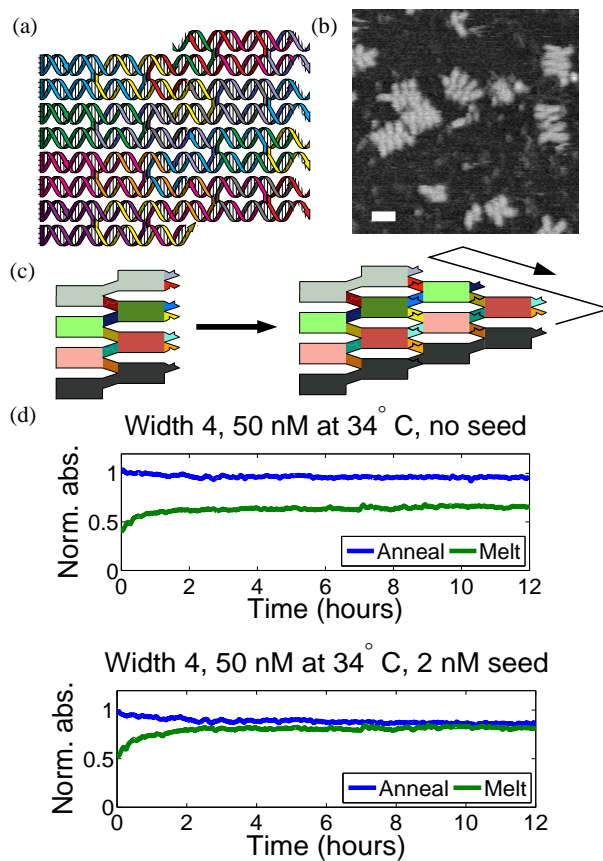


Figure 4.5: **Hysteresis in ribbon formation disappears when crystal seeds are present.** (a) A DNA structure designed to be a stable nucleus for ZZ4. The structure (the size of the helical regions and placement of crossover points) is identical to the structure of the tile lattice. (b) Atomic force microscopy image of the crystal seeds. The scale bar denotes 25 nm. Both intact and incomplete structures are seen. (c) Putative growth from a crystal seed: at every step, tiles can bind favorably (by two sticky ends) to produce the structure shown on the right, which can then grow through zig-zag growth. (d) Hysteresis of 50 nM ZZ4 with and without crystal seeds at 34 °C over 12 hours. The increased equilibrium absorbance in the sample with seeds is due to the added material.

Chapter 5

Copying Information Accurately by Using Proofreading

Abstract

Biological organisms are compelling demonstrations of the power of control over information transfer and computation at the molecular scale. In electronic computers, reliable computation is made possible by redundancy, which allows errors to be detected and corrected; increasing the amount of redundancy can exponentially reduce errors. Here, we examine experimentally whether using multiple levels of redundancy during self-assembly can analogously reduce assembly errors. We use DNA double crossover molecules to algorithmically self-assemble ribbon crystals which copy short sequences of information multiple times and measure error rates when each bit is encoded by 1 molecule, or redundantly encoded by 2, 3, or 4 molecules. Atomic force microscopy images of the products show that each additional level of redundancy decreases the assembly error rate by a factor of 3, with the 4-redundant encoding yielding an error rate less than 0.1%. While theory and simulation predict larger improvements in error rates than we observed experimentally, our experimental results suggest that by using sufficient redundancy it may be possible to cheaply and reliably self-assemble micron-sized objects with nanometer-scale detail.

5.1 Introduction

At the human-size scale, computation and information have transformed our world. Biology suggests that control over molecular information transfer and computation has the potential to do the same at the molecular scale. As examples, biological organisms have evolved the capacity for reliable, large scale self-assembly of objects like microtubules [DM97] or the ribosome [RSSF03], long signal-transduction cascades like those in animal development [Dav01] and tight control over metabolic processes [Fel92]. To engineer systems with the same capacities, it is necessary to develop design principles for chemical systems with the ability to reliably process and transmit information.

This paper concentrates on how to reliably transfer information during self-assembly. In an autonomous self-assembly process, all the information to guide assembly is encoded in the molecules themselves. Information transfer is not strictly necessary for self-assembly; if each molecule is only used once in a final product, as in protein folding or DNA origami [Rot06], self-assembly can occur reliably without multiple levels of information transfer. However, the requirement that each kind

of molecule be used only once is not practical for the assembly of micron-scale objects. Biology generally reuses components to assemble objects of this size, as evidenced by the self-assembly of networks of actin [PC86], microtubules [WBT68, FFLN77], or biomineralization [Man88]. In these cases, the molecules encode a pathway of assembly. Reliable information transfer along the pathway is necessary to produce a correctly formed final product.

Algorithmic self-assembly, a generalization of crystal growth, has been proposed as a general method of this kind for synthesizing supramolecular objects [Win96]. In algorithmic self-assembly, a set of tiles containing four binding sites with particular affinities executes a “program” for the assembly of an object. Such an assembly method is surprisingly powerful; abstractly, the assembly of such tiles (Wang tiles) into a lattice can simulate universal computation [Ber66], and in principle algorithmic self-assembly can be used to assemble arbitrary shapes, with the number of molecules required scaling only modestly faster than the Kolmogorov (algorithmic) complexity of the object [SW04]. Thus, algorithmic self-assembly is ideal for assembling large, complex objects that can be compactly described by a computer program such as defined-size shapes [RW00, ACGH01] or computer circuits [CRW04].

In practice, DNA double-crossover molecules [FS93, LYK⁺00, MSS99, YPF⁺03, HCL⁺05, CSK⁺04] with binding sites consisting of single stranded DNA segments have been used as tiles for algorithmic self-assembly (Figure 5.1a). Algorithmic self-assembly of DNA tiles has been demonstrated in one-dimension [Adl94, MLRS00a], and in two dimensions it has been used to produce complex, aperiodic patterns [RPW04, BRW05, BSRW07, FHP⁺07]. While these demonstrations indicate that the principles of algorithmic self-assembly are sound, the measured rate of assembly errors were high, between 0.1 and 10 percent. Since even a single assembly error can produce a malformed final product, such high error rates severely limit the applicability of this technique.

Error-free algorithmic self-assembly consists of a series of steps in which a tile binds to a growing crystal by at least two matching bonds. When several tiles can attach at different sites on the crystal, the order of assembly is non-deterministic, but it is straightforward to construct tile sets for which there is only one possible correct assembly product. Under slightly supersaturated conditions, such attachments are energetically favorable, but attachments of tiles to a crystal by fewer than two bonds are unfavorable. In practice, however, it is believed that transient unfavorable tile attachments occur often, and if a second tile attaches to it before it falls off, the mismatched tile can be “locked in,” resulting in an assembly mismatch error.

Proofreading tile sets work by interrupting the lock-in process. When an incorrect tile in a proofreading tile set attaches, further mismatches must occur within the tile’s block in order for the incorrect tile to become fixed in place (Figure 5.1b). Because sequences of multiple co-localized errors are relatively rare, assembly stalls, allowing the incorrect tile to fall off and correct assembly to proceed. The block size determines the number of mismatches that must occur before lock-in. Each additional required mismatch should, in theory, exponentially reduce the error rate [WB04]. It should therefore be possible to correctly assemble objects consisting of thousands to millions of tiles with only a modest increase in redundancy.

In this paper, we experimentally measure the amount by which the assembly error rate decreases as the proofreading block size is increased. We determine the mismatch error rate of 4 different tile sets, each of which copies a short, one-dimensional binary sequence with increasing amounts of redundancy. We show that, as theory predicts, increasing redundancy decreases the error rate by a multiplicative factor. However, our measurements indicate that the factor by which error rates are reduced is lower than what simulations or theory predict. Despite this limitation, these results

suggest that with sufficient redundancy, error-free algorithmic self-assembly of large objects should be feasible.

5.2 Experimental Design

Copying is a fundamental form of information transfer during self-assembly. Thus, to measure mismatch error rates, we designed a set of tiles that copy a short binary sequence over and over as they assemble a ribbon-shaped crystal. In our design, the correct attachment of a tile or block copies a bit in a sequence to a growing layer of the crystal, while the incorporation of a tile or block that matches only one binding domain incorrectly “flips” that bit in the sequence in the growing layer. Errors during copying are easy to spot experimentally, because an error in sequence copying is propagated when the sequence is recopied in the next layer.

We used the previously designed and characterized zig-zag tile set [SW07] as the basis of our design. A zig-zag ribbon crystal has a fixed width, in our case six tiles, but during assembly grows longer, adding one row of tiles at a time. Each tile in a zig-zag tile set is specific to a particular row in the crystal. The tiles in the middle four rows of the crystal encode either a “0” or a “1” bit to be copied. These tiles have four binding domains; the top left and bottom right domains encode the particular row that the tile may join. The top right and bottom left domains encode either a “0” or “1” bit. Under slightly supersaturated growth conditions, tiles prefer to attach to the crystal by two binding domains at exactly one location on each end of the crystal; each attachment creates a new binding site at either the row above or below the newly attached tile. In order for a tile to attach in a middle row, it must encode the correct row information *and* match the logical value of the tile in the previous row. Thus, while both “0” and “1” tiles may be present in solution, at each step the attachment of tile which matches the sequence of information encoded by the crystal is favored over the attachment of a tile which does not. The so-called double tiles in the top and bottom rows ends a series of tile attachments in one direction (from bottom to top or top to bottom) and creates a new binding site to initiate a new series of tile attachments in the opposite direction. The use of two alternating tile types in each of the middle row enforce the staggered placement of double tiles in top and bottom rows, ensuring that row by row, “zig-zag” growth can continue.

To measure how the rates of assembly error during copying decrease as the amount of redundancy increases, we designed 4 zig-zag tile sets which copy a sequence with increasing amounts of redundancy (Figures 5.1e-h). Because the same number of rows of tiles are available for copying in each tile set, the redundantly copied sequences contain fewer distinct bits: The 1-redundant tiles shown in Figure 5.1e copy a 4-bit sequence by using 1 row of tiles to store each bit of the sequence. By contrast, the 2-redundant tiles in Figure 5.1f copy a 2-bit sequence during their assembly—by using two rows to store each bit in the sequence. We used the 3- and 4-redundant tile sets in Figure 5.1g and 5.1h to measure the error rates when copying 1 bit stored by 3 and 4 rows of tiles respectively. We copied a second bit in the leftover row of the 3-redundant tile set.

We used a crystal seed to initiate the copying of a pre-determined pattern (Figure 5.2a) similar to a seed used previously to nucleate zig-zag ribbons [SW07]. The crystal seed is a stable structure that has binding domains for a particular sequence. Correct growth from the crystal seed first produces a “cone” of tiles that match sequence initiated by the seed (Figure 5.2b). Double tiles can then bind to the assembly and initiate zig-zag growth (Figure 5.2c).

5.3 Simulation

Based on what is known about the hybridization kinetics of DNA tiles [Rot00, Har04, SW07], we sought to qualitatively predict how error rates in our 4 tile sets decreased with increasing redundancy. We used the kinetic tile assembly model (kTAM) [Win98], which assumes that tile attachment occurs at a rate $k_f = 10^6/M/s$, approximately the measured attachment rate of short DNA oligos [Wet91], and that the ΔH° and ΔS° of attachment of a tile by two bonds is as previously measured for zig-zag ribbons [SW07]. While the energy of tile attachment by one bond has not been measured, we assumed it to be half the energy of attachment by two bonds. The kTAM uses stochastic kinetics [Gil76] to simulate each tile binding and unbinding event. Attachment or detachment of blocks of tiles, while possible in solution, were not modelled.

The error rate is dependent on two physical parameters, the concentration of tiles and the energy of attachment by two sticky end bonds. The rate of tile attachment at a given site on a crystal is $k_f[t] \equiv k_f e^{-G_{mc}}$ where k_f is the forward rate constant and $[t]$ is the free tile concentration. The rate of tile detachment by b sticky ends is $k_f e^{-\frac{b}{2}(\Delta H^\circ - T\Delta S^\circ)/RT} \equiv k_f e^{-bG_{se}}$ where ΔH° and ΔS° are the previously measured standard enthalpy and entropy of attachment by two bonds, respectively, T is absolute temperature, and R is the universal gas constant.

The kTAM predicts that algorithmic self-assembly is possible when tiles attach favorably only when they match at least two sticky ends on a growing ribbon. This occurs where $2G_{se} > G_{mc} > G_{se}$. For a given G_{mc} , assembly occurs most accurately just below the melting temperature, when $G_{mc} = 2G_{se} - \epsilon$ and becomes less accurate as G_{se} (and supersaturation) increases. The ratio $\tau \equiv \frac{G_{mc}}{G_{se}}$ is conventionally used as a measure of supersaturation for algorithmic self-assembly reactions [Win98, WB04, CG05].

To understand roughly how the error rate is a function of the physical parameters, we analytically estimated the assembly error rate for a given G_{mc} and G_{se} . In a 1-redundant tile set, a single mismatching tile can be locked in by the next tile, which can attach by two matching bonds. The probability that a mismatching tile is locked in before it can fall off, in terms of the rate constants in Figure 5.1c is $\frac{f}{f+r_1} \equiv \frac{e^{-G_{mc}}}{e^{-G_{mc}} + e^{-G_{se}}}$. The rate at which mismatched tiles are locked in is the rate at which such a tile attaches, $k_f e^{-G_{mc}}$ times this probability. (In this simple analysis, we ignore the probability that the tile that attached by two sticky end bonds will subsequently fall off.)

With a tile set that has $n > 1$ levels of redundancy, an incorrect attachment must be followed by another incorrect attachment which also tends to fall off (Figure 5.1d). For such a tile set, the rate of mistakes is therefore the rate of incorrect tile attachment times the probability that the next n tiles will both attach and lock in the first error, or roughly

$$k_f e^{-G_{mc}} \left(\frac{e^{-G_{mc}}}{e^{-G_{mc}} + e^{-G_{se}}} \right)^n. \quad (5.1)$$

This equation is approximate, as it neglects the probability that tiles will fall off and reattach during the lock-in process, or that, as above, the tile that attaches by two matching bonds will fall off. But roughly, each layer of redundancy is expected to reduce the error rate by a multiplicative factor $\frac{e^{-G_{mc}}}{e^{-G_{mc}} + e^{-G_{se}}} \approx e^{G_{mc}(1-\frac{1}{\tau})}$.

We measured the error rate during ribbon growth with kTAM simulations where $G_{mc} = 13$ across values of τ compatible with algorithmic self-assembly. (The value $G_{mc} = 13$ corresponds to a tile concentration of about 2.3 μM . Accurately measuring simulated error rates at the concentration used in experiments was computationally intractable.) The error rates predicted by

these simulations (Figure 5.2a) are consistent with Equation 5.1. For near-optimal values of τ , the simulations predict a decrease in error rate by a multiplicative factor with each additional layer of redundancy.

Because our analysis and simulations predict that the error rate will be highly dependent on the supersaturation, we assumed that the error rates in our experiments would also be highly dependent on the amount of supersaturation while the crystals are growing. In our experiments, crystals grew either from seeds or by spontaneously nucleating. We used simulations to determine when growth from seeds or spontaneous nucleation of new crystals would start new ribbon growth with the annealing schedule we used. In these simulations, G_{mc} was the concentration of tiles used in our experiments (50 nM for each tile) and $G_{se} = \frac{1}{2}(\Delta H^\circ - T\Delta S^\circ)/RT$ reflected previous experimental measurements of ΔH° and ΔS° . The temperature was reduced at the same rate as in our experiments. A crystal was considered nucleated when it consisted of at least 50 tiles. The results (Figure 5.3b) predict that there is no kinetic barrier to nucleation of crystals when seeds are present; as soon as growth of any kind is favorable, and even slightly before (due to fluctuations), crystals grow. In contrast, spontaneously nucleated crystals do not nucleate until the solution becomes fairly supersaturated and predicted error rates are much higher. We therefore predict that crystals that grow from seeds should contain fewer errors.

5.4 Results

Design. In our experiments, each logical tile in Figure 5.1 is implemented as DAO-E double crossover molecule [FS93] (Figure 5.1a). A DAO-E molecule consists of a set of short oligonucleotides that self-assemble into their pseudoknotted structures because of a preference for hybridization between Watson-Crick complementary subsequences. The “core” of a tile is double stranded and each end contains a 5 base pair single-stranded region, a sticky end which encodes the tile’s affinity for other tiles; each logical binding domain (Figure 5.1) is represented as a pair of complementary 5 base pair sticky ends. Double tiles have the structure of two single tiles that have been ligated. Where indicated in diagrams (Figure 5.1), double tiles do not have sticky ends. The cores of the “1” tiles contained two hairpins in the middle of one of its helices perpendicular to the plane of the lattice so that “1” tile could be differentiated from “0” tiles in atomic force microscope images [RPW04]. Tile sequences are designed as reported previously [SW07] using secondary structure minimization [See90, DLWP04] to prevent spurious interactions. By design, matching sticky ends have similar hybridization energies but pairs of ends that could mismatch have a hybridization energy of 0 according to the nearest neighbor model of DNA hybridization [BCT00].

The crystal seed contains a crossover structure identical to that in DAO-E lattice [WLWS98] but its strands are woven so that the structure cannot fall apart without breaking at least 16 base pairs, in contrast to the 5–10 base pairs that must be broken for a tile to detach from the outside of a DAO-E tile crystal. The melting temperature of a similar seed structure at the concentration we use here has been measured as 62 °C, well above the melting temperature of the ribbons (about 34 °C at 50 nM [SW07]). The crystal seed can be used nucleate any desired sequence by changing the strands that create the sticky ends along the right edge.

Tile and seed assembly. To test that the tiles formed properly, we assembled individual tiles by annealing their component strands from 90 °C to 20 °C at 1 °C per minute. (This and all future experiments were performed in TAE buffer with 12.5 mM added $MgCl_2$.) Polyacrylamide gel electrophoresis showed that each tile assembled into a single product with at least 80% yield. We

annealed the crystal seed structure in the same manner. Atomic force microscopy showed structures that look like the designed structure (Figure 5.4a).

To test that 6 tile wide ribbons formed, we annealed the strands for the tiles of the non-proofreading ribbon from 90 °C to 40 °C at 1 °C/hour to allow tiles to form and then in the previously determined regime of tile supersaturation [SW07], 40 °C to 20 °C, at 1 °C/hour. Atomic force microscopy revealed that each ribbon copied a pattern with occasional errors (Figure 5.4b). Each of the 10 distinguishable patterns (because the top and bottom tiles the zig-zag crystal are indistinguishable with AFM, the orientation of asymmetric patterns such as 1000 can not be determined) were seen. In experiments described later, all possible patterns with 1, 2, and 3 levels of redundancy were also observed, indicating that all tiles correctly performed their logical function.

Ideally, assemblies with all possible patterns would nucleate and grow at the same rate (Figure 6(a), blue bars). However, the formation temperature of DNA tile lattices without hairpins has been measured to be slightly higher than the formation temperature of DNA tile lattices with hairpins [Bar07]. As a result, when the two tile types are mixed, lattices with all “0” tiles form first, resulting in an excess of these lattices, and preventing the “0” tiles from being used in lattices encoding other patterns. To determine whether the assembly of some patterns was preferred over others, we tabulated which patterns were copied with 25 1 μm square images at random locations on the surface. Patterns which contained all or almost all 0 or 1 tiles were seen more frequently than would be expected, while those with a mixture of 0s and 1s were rare (Figure 5.6a, green bars). We inferred that the formation temperature of ribbons containing all “0” tiles was, as seen previously with other DNA crystals, slightly higher than the formation temperature of other ribbons. This effect made the concentration of “0” and “1” tiles unequal during annealing, affecting the mismatch error rates.

To address this problem, we repeated the experiment but added two kinds of crystal seeds which nucleated the patterns 0101 and 1010 respectively to a total concentration of 2 nM, as simulations indicate that ribbons would grow first from seeds. If most crystals grew from seeds and both kinds grew at the same rate, tiles would be used up equally, permitting an unbiased measure of mismatch error rates. Seeds were added at 50 °C, above the melting temperature of ribbons, but below the melting temperature of the seeds. In this experiment (Figure 5.6a, red bars), the pattern 0000 was still preferred, but the seeded patterns were present in more than 20% of ribbons, 10 times more frequently than when the seeds were not used.

Error rates. To determine the rate at which assembly errors occurred in the 1-redundant tile set, we counted the number of bits that were correctly and incorrectly copied in about 65 1 μm square images of ribbons annealed with crystal seeds (about 25,000 bits total). The measured error rate was approximately 0.01 (Figure 5.4b, bar 1). Crystals with seeded pattern should nucleate first (see Figure 5.3b), and therefore grow under less supersaturated conditions where the error rate is lower than spontaneously nucleated crystals. As expected, the error rate in crystals with the seeded pattern (Figure 5.4b, bar 6) was slightly smaller than the error rate measured by considering all crystals.

To determine how error rates decrease when bits are redundantly encoded, we repeated the above experiment with the 2-, 3-, and 4-redundant tile sets. The crystal seeds nucleated the patterns 01 and 10 with the 2- and 3- redundant tile sets, and the patterns 0 and 1 with the 4-redundant tile set. The results (Figure 5.6b, bars 2-4) show that with each level of redundancy that is added, the error rate decreased by a factor of about 3, with the error rate for 4-redundant tile set being less than 0.001. Surprisingly, the error rate in fourth row of the 3-redundant ribbons, was about three

times higher than the error rate in copying the bits in the 1-redundant tile set (Figure 5.6b, bar 5).

5.5 Discussion

The observed error rates while copying information in zig-zag ribbon crystals agree with the qualitative results predicted by theory and by simulation—with each level of redundancy, we see an approximately multiplicative reduction in the error rate. Additionally, we see that the biggest reduction in error rates comes with introducing the first level of redundancy—we observed a factor of about 4 improvement in this case and slightly smaller improvements as the amount of redundancy increases. This pattern is also in qualitative agreement with the simulations.

However, simulations and theory predict a much lower absolute error rate than we observed. These discrepancies are of several orders of magnitude in some cases—understanding the reason for them seems therefore to be of central importance. One obvious possibility is that our model, kTAM, is inaccurate because we used the wrong parameters. Parameters such as the rate of tile attachment and the entropy lost due to hybridization are inferred from studies of small oligos rather than from studies of DNA tiles [Win98]. Additionally, it has never been confirmed that the energy of attachment of a tile by a single sticky end is half the energy of attachment by two sticky ends. In other studies [CSGW07], relaxing the last assumption produced simulated error rates more in line with what was measured. In other work, a smaller forward rate produced results that were more in line with experimental measurements [SW07]. Additionally, the kTAM does not consider experimental non-idealities such as stoichiometry errors between strands and different formation rates of tiles.

Experimental non-idealities likely also contributed to the high error rate. The crystal seeds we used were not as effective at nucleating ribbons as larger, more sturdy DNA crossover molecules used previously [BSRW07]. Since simulations suggest that spontaneous crystal nucleation occurred at high supersaturation, some growth doubtless occurred in a regime where error rates are comparatively high. One other factor that may have produced errors is ribbon joining, which has been observed at relatively high rates at room temperature [SW07]. It was not possible for us to separate errors that resulted from joining from those that resulted from a series of single-tile mismatches. Reliable nucleation of ribbons from a seed would prevent joining and would therefore allow a more accurate measurement of ribbon growth rates.

Despite the comparatively large observed error rate and comparatively small observed error rate reductions due to redundancy, we did not observe a limit on the improvement that can be obtained with redundancy; the introduction of each additional layer significantly reduced error rates. Thus, it now seems possible, without any improvement in experimental technique, to assemble ribbons with virtually no errors simply by using enough redundancy. While large amounts of redundancy increase the size of the product, it should be possible to mitigate this increase by using either tile sets which introduce redundancy without increasing size [FM04, RSY05], or by using smaller tiles, possibly consisting of a single strand [YHS⁺07].

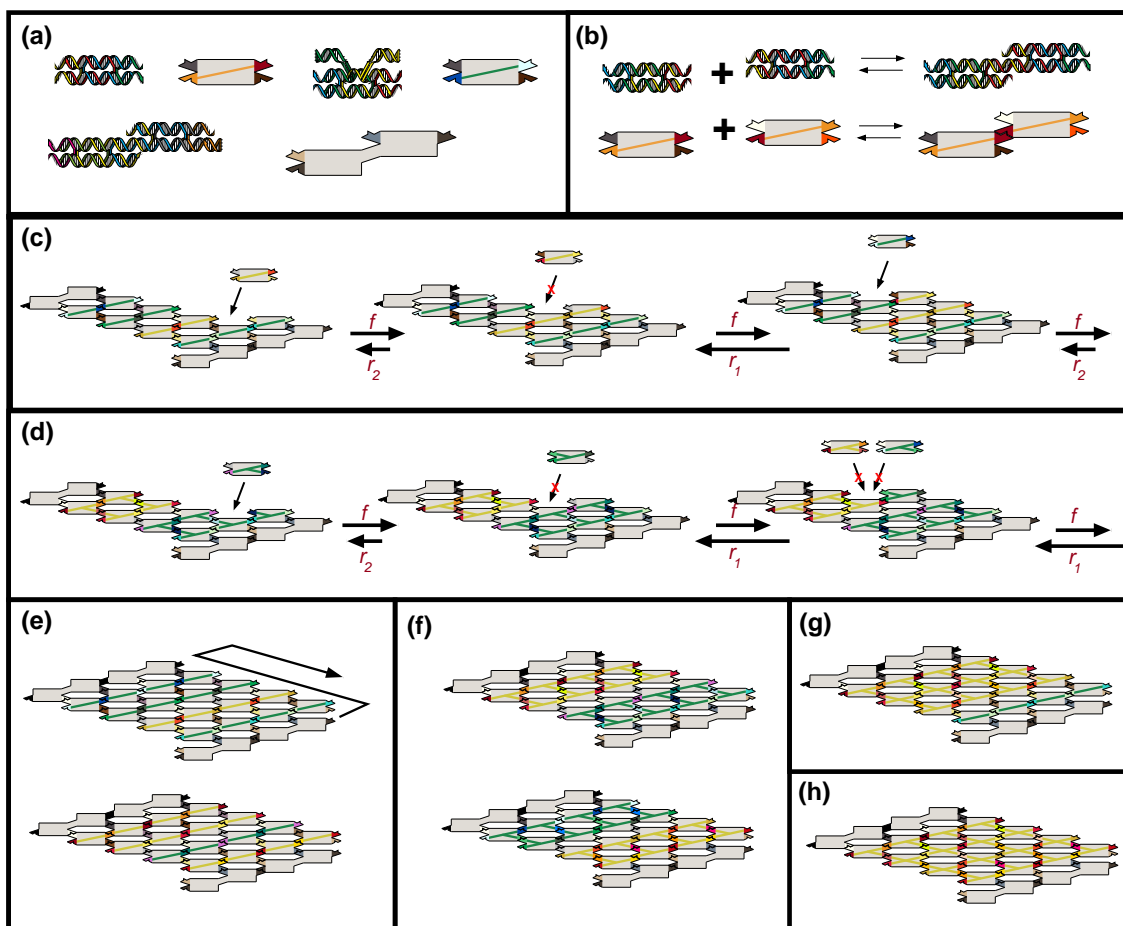


Figure 5.1: **DNA tiles for ribbons that copy information.** (a) The DNA strands that comprise the three kinds DAO-E double crossover molecules used in this paper, with their respective composite diagrams. Each molecule consists of 4–6 synthetic DNA strands which self-assemble into the products shown here because of a preference for Watson-Crick complementarity. (b) Tiles bind by hybridization of their sticky ends. Non-complementary sticky-ends tend not to hybridize. (c) Tiles favorably attach to a crystal by two bonds (left). In the tile set shown, tiles that do not propagate the existing sequence (colored) stripes can make at most one bond. These tiles can attach transiently but generally fall off quickly (center). If another tile attaches before a mismatching tile falls off (right), the tile is locked in, causing an assembly error. (d) Proofreading reduces the mismatch error rate by interrupting the lock-in process. Here, each bit is copied by two rows of tiles instead of one. Yellow and green lines (representing “0” and “1” respectively) connect tiles that copy the same bit. When a tile that mismatches attaches (middle), because the bonds within the two tile block are unique to the block, no tile matches more than one bond at the new growth site. Thus, another mismatch must occur before the tile can be locked in (right). (e) A set of tiles (assembled into two ribbons) that copy a 4 bit sequence. While ribbons copying only two sequences are shown, all 16 possible 4 bit sequences can be propagated by ribbon growth. (f)–(h) Tile sets in ribbon form that, respectively, copy sequences 2-, 3-, and 4-redundantly. Yellow and green lines as in (d). The tile set in (g) copies 1 bit with three rows and 1 bit with the 1 remaining row. As in (e), for each tile set, all possible binary sequences can be copied. The tiles to propagate a green bit in three rows (3-redundantly) and a yellow bit in one row (1-redundantly) (g) and the tiles to copy a green bit with four rows in the tile set in (h) are not depicted.

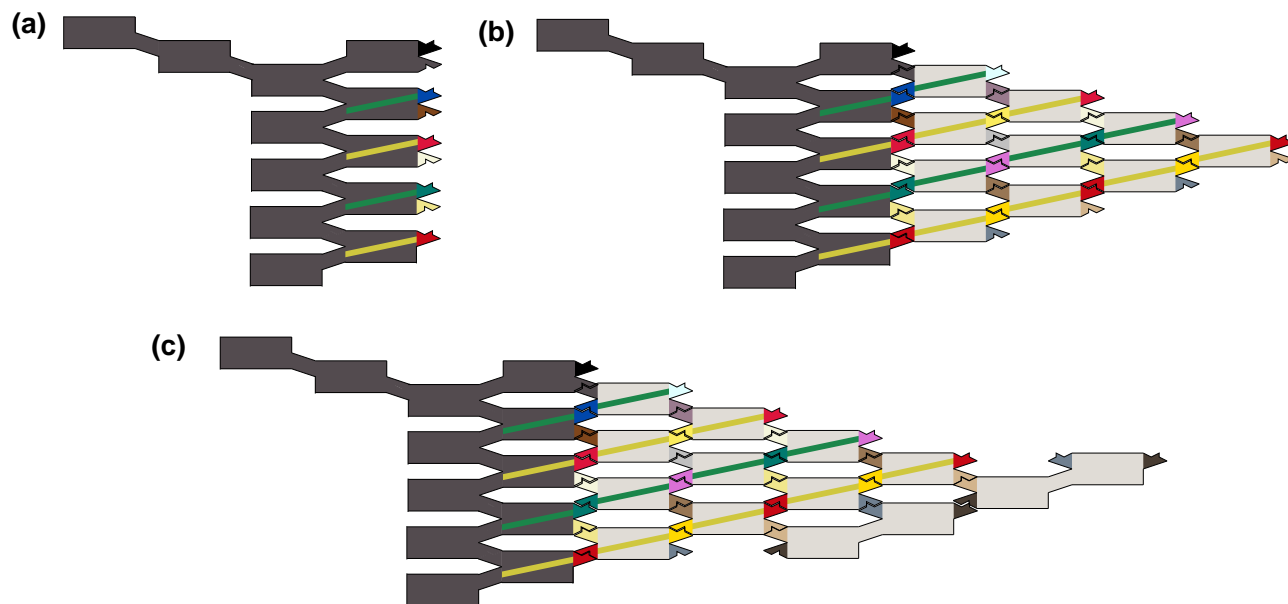


Figure 5.2: **Nucleation of a ribbon from a crystal seed.** (a) The growth of the tiles in Figure 5.1e on a crystal seed structure, shown in composite by the gray structure. The binding sites on the seed are such that this seed propagates the illustrated green-yellow-green-yellow sequence. (b) Initial valid growth off the seed produces a V-shaped assembly of tiles. While the assembly order is non-deterministic, all possible assembly orders correctly copy the sequence. (c) After enough layers of tiles have accumulated, a double tile can attach by two bonds to the bottom edge. The attachment of a second adjacent double tile allows zig-zag growth (see arrow, Figure 5.1e to commence).

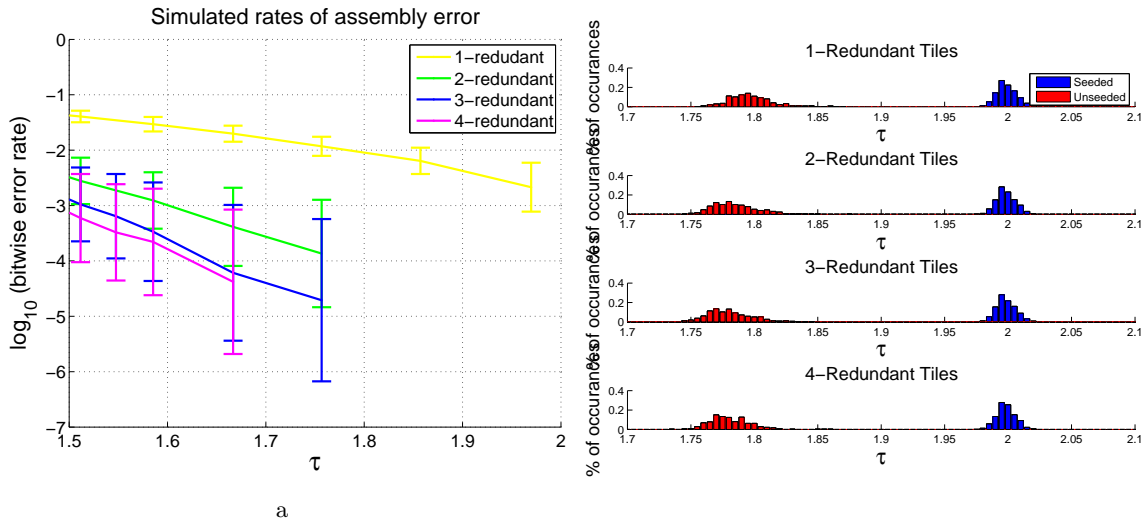


Figure 5.3: **Simulated assembly error rates and nucleation temperatures.** (a) Simulated error rates during ribbon assembly. Tile concentration is $\sim 2 \mu\text{M}$ at various τ , a measure of supersaturation (See text; $\tau = 2$ is the melting temperature of the ribbons.) Error rate measurements were made over the growth of a 500 row long ribbon. The free tile concentration was held constant over the course of the simulation. Error bars indicate 2 standard deviations. (b) Histogram of τ values at which crystals reach a certain size in a series of stochastic kinetic simulations of annealing ribbon tiles. In the simulation, a 10^{-17} L drop containing 50 nM of each tile was annealed from 40 °C to 20 °C with the temperature decreasing 1 °C per hour. At each temperature, the sticky end energy was set using the previously measured ΔH° and ΔS° of tiles attaching by 2 sticky ends. Plotted here are the supersaturations when a crystal reached 80 tiles. Free tile concentration was held constant over the course of the simulation. Each histogram comprises at least 100 simulations.

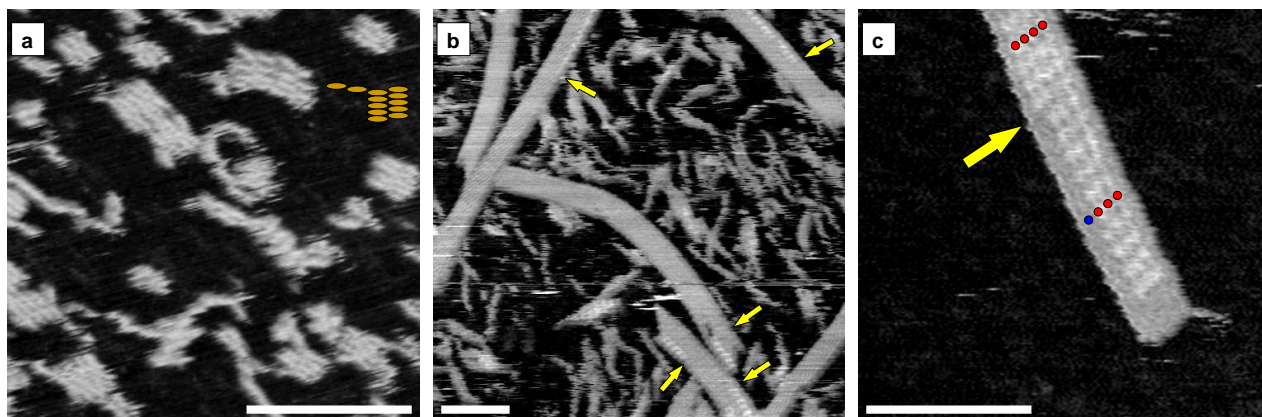


Figure 5.4: **AFM images of seeds and ribbons.** Scale bars represent 100 nm. Yellow arrows indicate locations of assembly errors. **(a)** AFM image of crystal seeds, which have the structure of 12 ligated DAO-E tiles as shown in Figure 5.2a and by the brown illustration. Seeds formed with 40%–50% yield. **(b)** AFM image of 1-redundant ribbons grown without seeds. Images portrayed either fully formed ribbons of the correct tile width or much smaller detritus, as is seen here in the background. Brighter spots are one tiles, which have hairpin attachments in their cores and therefore more AFM contrast. **(c)** A seed-nucleated 3-redundant ribbon. Blue and red dots label the “1” and “0” elements of a sequence respectively. The “tail” off the end of the ribbon is the extra DNA added to the edge of the seed so that it can be differentiated from ribbons. On most ribbons with the nucleated pattern, this tail was not visible.

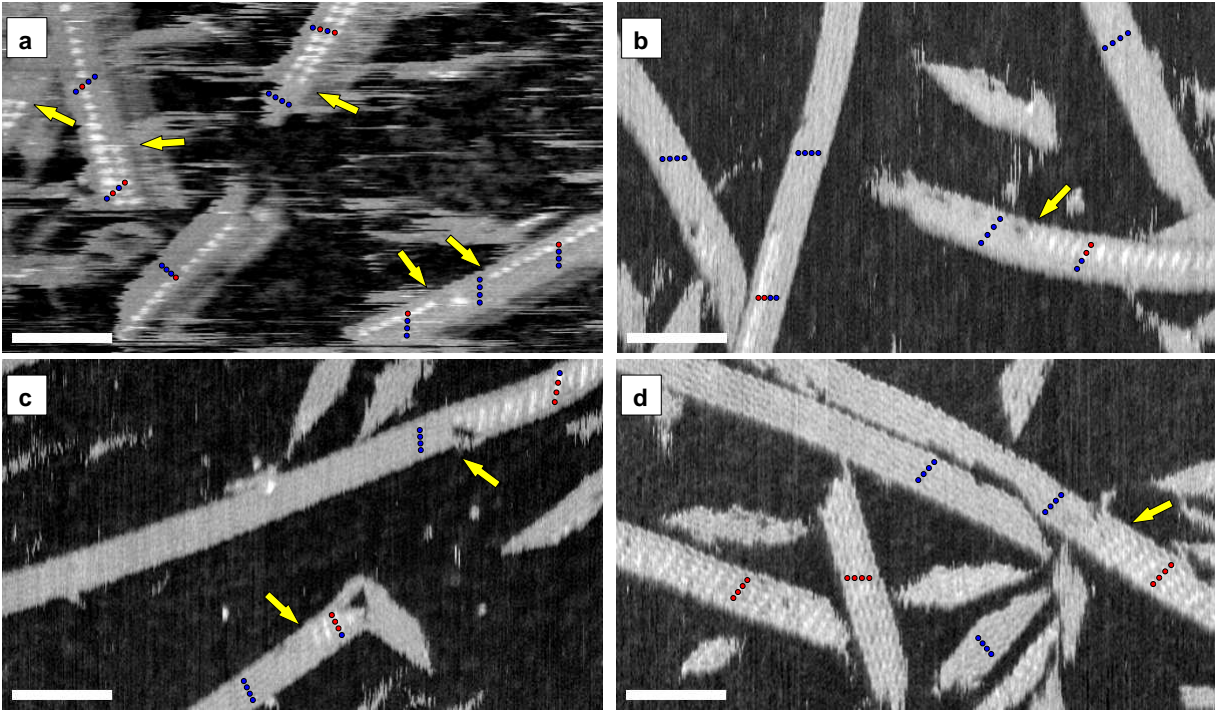


Figure 5.5: **Sample AFM images of ribbons copying sequences.** Scale bars and yellow arrows as in Figure 5.4. Blue and red dots label the “1” and “0” elements of a sequence respectively. (a) 1-redundant ribbons, (b) 2-redundant ribbons, (c) 3-redundant ribbons, (d) 4-redundant ribbons

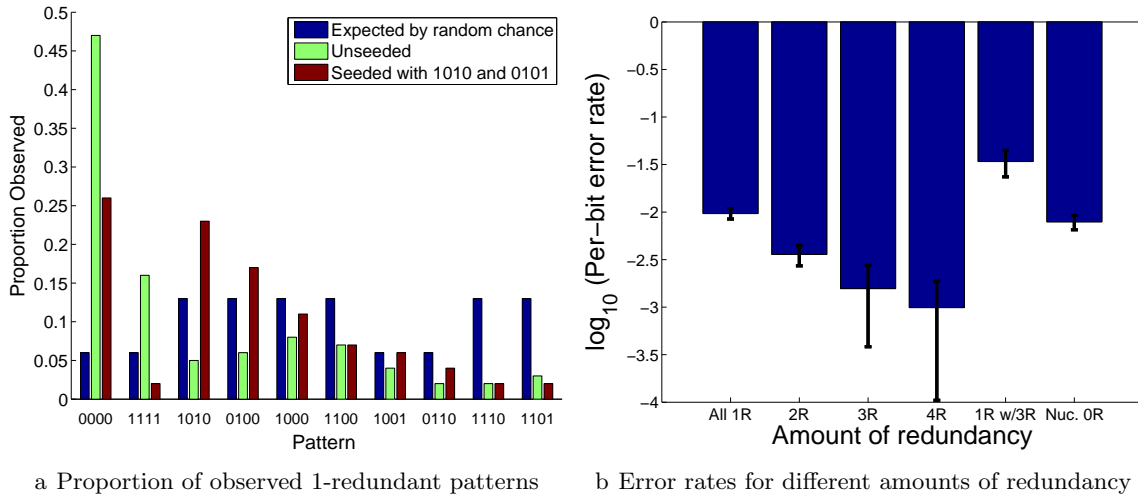


Figure 5.6: **Experimental Results.** (a) A tabulation of how often ribbons with each possible pattern appeared in experiments, based on 25 images in each case. A uniform distribution of pattern sequences is given by the blue bars; patterns were distributed randomly. Without seeds, the 0000 pattern is seen much more frequently than other patterns (green bars), possibly because ribbons without hairpin markers have a higher formation temperature and therefore nucleate first, using up the “0” tiles. When seeds that nucleate the 0101 pattern (red bars) are present, this pattern is enriched, but the 0000 is still seen most frequently. (b) Per-bit error rates for 1, 2, 3, and 4-redundant tile sets (bars 1–4). Surprisingly, the error rate of the non-redundantly encoded extra bit in the 3-redundant tile set had a much higher error rate than any of the bits in the 1-R ribbon (bar 5). The error rate in 1-redundant ribbons with the nucleated patterns is less than the error rate of all 1-redundant ribbons (bar 6), supporting the hypothesis that these ribbons nucleated and assembled on average in less supersaturated conditions.

Part III

Evolution

There are some enterprises in which a careful disorderliness is the true method.

Herman Melville, *Moby Dick*

Chapter 6

The Potential for Evolution of Complex Programs

Abstract

An enduring mystery in biology is how a physical entity simple enough to have arisen spontaneously could have evolved into the complex life seen on Earth today. Cairns-Smith has proposed that life might have originated in clays which stored genomes consisting of an arrangement of crystal monomers that was replicated during growth. While a clay genome is simple enough to have conceivably arisen spontaneously, it is not obvious how it might have produced more complex forms as a result of evolution. Here, we examine this possibility in the tile assembly model, a generalized model of crystal growth that has been used to study the self-assembly of DNA tiles. We describe hypothetical crystals for which evolution of complex crystal sequences is driven by the scarceness of resources required for growth. We show how, under certain circumstances, crystal growth that performs computation can predict which resources are abundant. In such cases, crystals executing programs that make these predictions most accurately will grow fastest. Since crystals can perform universal computation, the complexity of computation that can be used to optimize growth is unbounded. To the extent that lessons derived from the tile assembly model might be applicable to mineral crystals, our results suggest that resource scarcity could conceivably have provided the evolutionary pressures necessary to produce complex clay genomes that sense and respond to changes in their environment.

6.1 Introduction

Developments in DNA computing have shown that computation can be embedded in crystal growth processes [Win96, RPW04], with potential applications to combinatorial search problems [LL00, MLRS00b] and to fabrication tasks in nanotechnology [CRW04, BRW05]. But does crystal computation have any relevance to what we observe in nature? We speculate here about a possible connection to the origin of life.

The background for this argument is a hypothesis, proposed and developed by Graham Cairns-Smith [CS66], that the first primitive “organisms” were clay crystals. In this theory, information (the first “genes”) consisted of patterns stored as variations in crystal structure that could be propagated during crystal growth. For example, in some layered silicate clays, there are two

distinct layer types that appear in a cross-section as a sequence that could be considered the crystal’s genotype. Replication occurs by periodic physically-induced fragmentation of crystals into smaller pieces containing the same genotype, leading to exponential increase in the number of organisms. Cairns-Smith considered several types of selective pressures that could have been present and would have resulted in favoring the growth of crystals with non-trivial genotypes. For example, the structure of a layer sequence could result in different crystal morphologies and different susceptibilities to fragmentation. He further envisioned the clays interacting with organic molecules, somehow leading to novel clay-produced organic molecules and eventually to the genetic takeover of organic life forms [CS82].

One of the strengths of Cairns-Smith’s hypothesis is that it is easy to see how Darwinian evolution could have gotten started by geological processes. However, while clays are known to interact with organic molecules [PEG⁺95, HFS03], it is hard to know whether these interactions could have provided evolutionary pressure toward increasing complexity. It is therefore interesting to ask whether there are other possible mechanisms that could have stimulated the original evolution of complex sequence information. In fact, it is hard to envision how anything so simple that it could have arisen spontaneously could have evolved into the remarkable complexity of form and function found in modern biological organisms. The capacity of evolutionary systems to create increasingly complex forms with increasingly adapted function—so-called open-ended evolution—remains poorly understood. To our knowledge, there are no examples in artificial life [BMP⁺00], *in vitro* chemical evolution [WJ97, Joy04], or *in vivo* directed evolution [YWA02] that have convincingly demonstrated open-ended evolution.

The conceptual and physical simplicity of crystal evolution makes it a promising place to examine these issues. In this paper, we consider whether there are properties of crystal growth that can lead to open-ended evolution. Examining the capacity for interesting evolutionary landscapes requires distinguishing between crystal genotype and phenotype—what is the genetically-determined function performed by crystals that gives rise to a selective advantage? In this paper, we investigate the ability of crystals to respond to selection pressures using computations that occur during growth.

The notion that a crystal can perform a computation is based on the observation that the tiling problem (the question of whether a set of geometric shapes can tile the plane) is undecidable: any problem solvable by a Turing machine can be expressed as a question of whether a particular set of shapes can tile the plane [Wan62]. This observation led to a constructive method of computing by arranging tiles into a lattice [Win96]. These tiles can be viewed as analogues for crystal monomers; attachment at a specific site in a growing crystal is determined by how well a tile’s shape fits in the growth site. The binding of a particular tile at a particular site can be viewed as a computational or information transfer step.

It is reasonable to assume that crystals grow in environments where their monomers are present in solution at different (possibly time-varying) concentrations. In such an environment, which crystal patterns grow the fastest? Are there environments in which crystals must perform computations in order to grow quickly? We show that crystals can sense the environment and respond by making use of the most abundant tiles. We call this feature a “crystal metabolism” because the computation they perform controls the extraction of resources from the environment. In particular, we are interested in whether there are environments that lead to the evolution of increasingly complex crystal metabolisms.

We examine these issues in principle within a previously described tile assembly model [RW00],

which is a generalized crystal growth model that has been used to study algorithmic self-assembly of synthetic DNA tiles [WLWS98, RPW04, BRW05]. It has been previously argued [SW05b] that DNA tile crystals have the capacity for Darwinian evolution of the sort imagined by Cairns-Smith. Insights derived from studying this model should be applicable to crystals composed of a variety of other materials, such as proteins [FFS⁺95, CDVW04], RNA [CSK⁺04], or even macroscopic tiles [BTCW97, Rot00] or clay minerals [CSH86].

The tile assembly model describes crystal monomers as square or rectangular tiles with each unit edge labeled to indicate how it fits with other monomers. Crystal growth proceeds by accretion, with single tiles being added to the crystal at sites where they make a sufficient number of contacts. In this paper we consider a version of this model in which (a) a tile may be added to a site if labels on at least two edges match those presented by the crystal at that site, and (b) monomer tiles arrive at potential binding sites with a frequency proportional to their concentration in solution. Occasional violations of rule (a) are referred to as “mutations.” A system consists of a set of tiles, along with the concentrations of those tiles in solution. Because of the particular choice of matching rules, each tile set implicitly determines what arrangements of tiles can grow as crystals. If certain arrangements grow faster than others under given conditions, then we consider these arrangements more fit. Later we will discuss how growth rates relate to the rate of exponential reproduction.

To discuss evolution, it is helpful to consider three aspects of an evolutionary process. First, a self-replicating entity carries information which directs behavior. The space of achievable behaviors is therefore inherently dependent on the material from which they are constructed. Second, there must be an environment in which certain behaviors have selective advantage; this provides the stimulus for evolution to discover complex solutions. Third, for evolution to proceed quickly there must be a route via mutations in which ever more complex behavior is achieved by a series of incremental steps. In crystal evolution, the first aspect (potential for complexity) derives from a choice of a particular tile set; this determines the behavior implicit in the crystal growth. The second aspect (stimulus for complexity) derives from the conditions in which the crystals are grown, e.g., tile concentrations and temperature, etc. The third aspect (the route to complexity) is difficult to predict. We will not address it here; for our purposes, it is enough to know that the fittest crystals could arise through (possibly extremely unlikely) mutations or spontaneous generation. However, understanding the route to complexity is an important goal for future studies of crystal evolution.

In addition to the above, open-ended evolution seems to require that the selective pressure provided by the environment is distinct from the potential for evolution, in the sense that different environments must lead to distinct functional solutions. For example, in modern biology, the universal potential of biochemistry—the ability of DNA to code for proteins and other macromolecules that create seemingly arbitrarily sophisticated and complex machines—makes it possible for living organisms to adapt to an incredible variety of environmental niches, opening the way to ever-increasing complexity. Thus, searching for tile sets capable of open-ended evolution, we first see that a tile set plays the role of a “chemistry” in the sense that it defines the rules by which tile-based “organisms” can grow and function, and we further expect that we are looking for a tile set that displays some sort of universality of behavior.

Following this intuition, we argue that tile-based crystals can exploit computation to enhance their growth rate, and that this can lead to evolutionary processes resulting in increasing complexity of crystal structure. We introduce a framework for studying the evolution of metabolic control in crystals under resource-limited growth conditions.

Within this framework, we design tile sets and environments that give rise to evolutionary land-

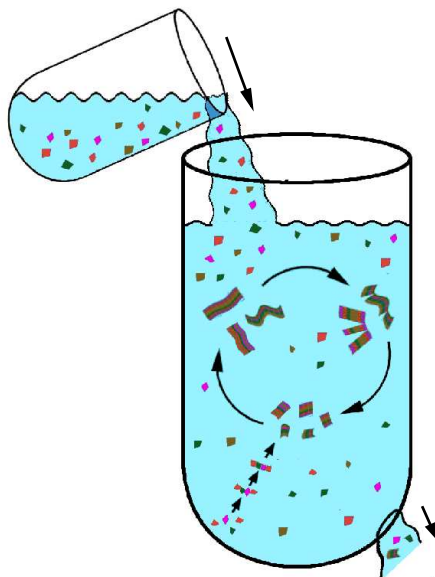


Figure 6.1: **The zig-zag crystal life cycle.** Zig-zag crystals grow by copying their sequences of DNA tiles. Reproduction occurs when a crystal is broken by external mechanical force (e.g., shearing, as shown in the upper right of the test tube). The small crystals that are the result of division (center) continue growing, and eventually become large enough (top left) to split again. The materials required for growth (tiles) are constantly replenished by an inward flow, while an outward flow removes slow-growing crystals from the population. Occasional mutations are propagated during growth and are eventually replicated. Similarly, new assemblies are occasionally generated spontaneously (lower left) from single tiles. Once they reach a certain size, these spontaneously generated assemblies can also grow and reproduce.

scapes in which crystals that perform more effective computations are fitter. We provide two main examples. First, in order to provide simple examples of metabolic evolution, we describe a tile set that can encode programmable logical computations. Second, to emphasize that arbitrarily complex computations can in principle be used by crystals to sense and respond to their environment, we exhibit a tile set capable of universal computation.

6.2 Zig-Zag Ribbon Evolution

Previously, it was suggested that DNA tile assemblies could in principle evolve through cycles of crystal growth and splitting [SW05b] (Figure 6.1). The zig-zag tile set described in that work produces ribbon-like crystals that copy information along the length of the crystal. The tile set includes a group of square tiles, and two rectangular tiles called *double tiles*. Logical representations of the tiles that comprise a basic zig-zag ribbon are shown in Figure 6.2a.

When growth occurs according to the tile assembly model, tiles are added to a ribbon in a zig-zag pattern shown in Figure 6.2b. Only tiles that match at least two edges can attach. Given this constraint, the design of the tiles is such that at any moment there is just one tile that may be added to each end of the ribbon. The addition of each new row can be viewed as the copying of the

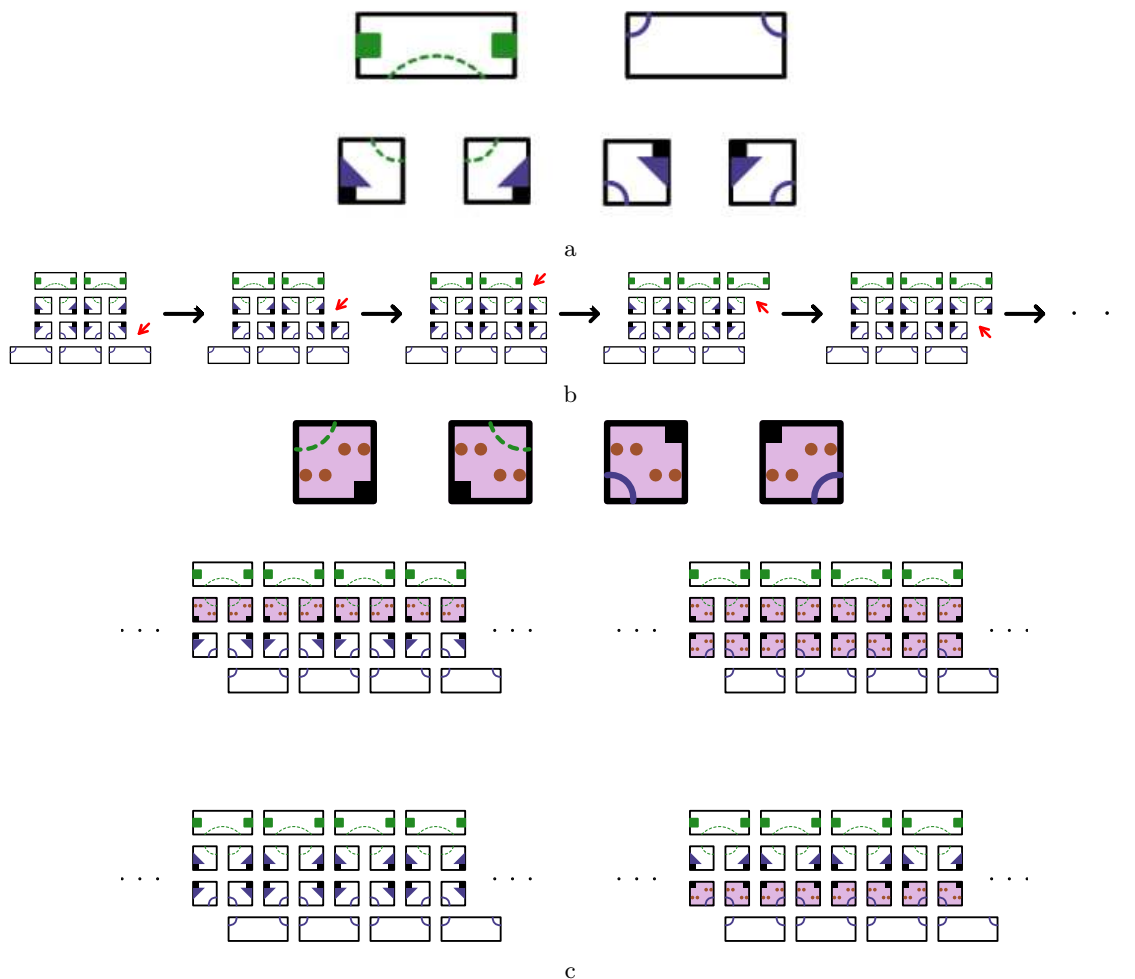


Figure 6.2: **The zig-zag tile set.** (a) The basic width-4 zig-zag tile set consists of six tile types. Tiles cannot be rotated. The tiles shown here have unique bonds that determine where they fit in the assembly: each label has exactly one match on another tile type. (b) The zig-zag tiles are designed to form the assembly shown here. Tiles can attach to the crystal where they match two edges of the crystal. Two alternating tiles in each column enforce the placement of the double tiles on the top and bottom, ensuring that growth continues in a zig-zag pattern. While growth on the right end of the molecule is shown here, growth occurs simultaneously on both ends of the molecule. At each step, a new tile may be added at the location designated by the small arrow. (c) The tile set shown in Figure 6.2a forms only one kind of assembly. The addition of the four tiles shown here allows four types of assemblies to be formed. Each assembly grows by copying its sequence (the vertical cross-section of tiles).

information in the previous row. Using the tile set shown in Figure 6.2a, this copying is trivial—only one sequence type is possible. However, the requirement that a tile attach by two bonds means that it must match both its vertical neighbor, either above and below, and its horizontal neighbor, to the left or right. In the case that several tile types may match the vertical neighbor, as shown in

Figure 6.2c, only the tile type that already appears in the row will match the horizontal neighbor. Only this tile can be added, so that information is inherited from layer to layer. As an example, the tile set shown in Figure 6.2c has two tiles for each position and therefore can propagate one of four strings. This construction can be generalized to an additional number of bits by adding tiles to the tile set in Figure 6.2c— 2^n sequences can be propagated by a tile set containing $4n + 2$ tiles. Further, a related tile set containing only 6 tiles can copy binary sequences of arbitrary width.

The growth of a crystal increases the number of copies of the original information present in the ribbon but does not produce any new growth fronts, so copying occurs at a constant rate. This information copying can be accelerated by periodic forces that cause ribbons to break. For each new ribbon that is created by breakage, two new growth fronts become available. Repeated fragmentation will therefore exponentially amplify an initial piece of information. Occasionally, a tile matching only one bond rather than two will join the assembly, resulting in a copying error, which will also be inherited. Such copying errors, inevitable in any physical implementation of tiles (e.g. [Win98]), will lead to evolution if ribbons with certain sequences grow faster than others.

6.3 Dynamics of Crystal Evolution

In this section we formulate a simple dynamical model to determine whether crystal growth and breakage leads to selective amplification, and to elucidate which properties of the environment and tile set are important in determining the replication rate of a sequence. The model tracks the concentration of crystals of each possible sequence. For a sequence s , two parameters are of interest: F_s , the number of growth fronts that can copy sequence s , and R_s , the number of columns of tiles, totaled over all crystals, with sequence s . The number of columns, R_s , increases at a rate proportional to the number of growth fronts, F_s , times the rate at which a new column can be added to a growth front, k_s . As this model is meant to be used in cases where growth rates depend upon the sequences s , k_s depends on s , reflecting the influence of tile set and environment. New growth fronts are produced when assemblies split into two pieces. We'll assume that splitting occurs with equal probability at each column, at a splitting rate p_s per column. (Again, this might be sequence-dependent.) Crystals die at rate f by being flushed out of solution.

Assuming that the growth rate is greater than the splitting rate ($k_s > p_s$), the dynamics of this system can be described by two linear differential equations for each sequence.

$$\frac{d}{dt} \begin{bmatrix} R_s \\ F_s \end{bmatrix} = \begin{bmatrix} -f & k_s \\ p_s & -f \end{bmatrix} \begin{bmatrix} R_s \\ F_s \end{bmatrix} \quad (6.1)$$

Because mutations are not included in this model, pairs of equations describing the growth and replication of a sequence are decoupled from equations describing the dynamics of other sequences. It is therefore not difficult to solve them; for a given sequence, the eigenvalues of the solution are $-f \pm \sqrt{k_s p_s}$. Assuming the death rates for all sequences are the same, sequences with faster growth rates and higher splitting rates are therefore amplified more than sequences that grow and split more slowly. As the death rate increases, only the sequences for which $\sqrt{k_s p_s}$ is large remain in solution. These sequences are selected for.

How does crystal evolution compare to RNA or DNA replication, e.g., of viral or bacterial genomes [Eig71], in which the growth rate is proportional to the concentration of sequences? As the decaying eigenmode dies away, $R_s \rightarrow \sqrt{\frac{k_s}{p_s}} F_s$. (The ratio $\sqrt{\frac{k_s}{p_s}}$ can be interpreted as half the average length of growing and splitting crystals. The ratio is halved because each crystal has R_s

rows and two growth fronts.) A new equation that measures only the dynamics of F_s , assuming $R_s \approx \sqrt{\frac{k_s}{p_s}} F_s$, is

$$\frac{d}{dt} F_s = \sqrt{k_s p_s} F_s - f F_s. \quad (6.2)$$

With the addition of mutation, where a mutation from sequence s to sequence t happens at rate m_{st} , this model is equivalent to Manfred Eigen’s model of RNA replication [Eig71] with the replication rate of a sequence replaced by the parameter $\sqrt{k_s p_s}$:

$$\frac{d}{dt} F_s = \left(\sqrt{k_s p_s} - f \right) F_s + \sum_t (m_{ts} F_t - m_{st} F_s). \quad (6.3)$$

Thus, this model is a generalized version of Eigen’s model, with DNA or RNA replication being the special case where $k_s = p_s$. When $k_s > p_s$, the fitness ($\sqrt{k_s p_s}$) of a crystal in a given environment depends on both the growth rate and the splitting rate. Thus, to show that a particular sequence s is fit, we must therefore show that $\sqrt{k_s p_s}$ is large, and that for unfit sequences t , $\sqrt{k_t p_t}$ is small.

6.4 A Zig-Zag Ribbon Metabolism

By what basis might some zig-zag crystals grow faster than others? In some models of crystal growth [Win98], the rate of attachment of a tile to a crystal is proportional to the concentration of the tile in solution, and the rate at which a tile is removed is related to the energy loss due to breaking the bonds between the tile and the rest of the crystal. Thus, the growth rate of crystals can be made faster by increasing the concentration of their component tiles in solution. Increasing growth rates increases a crystal’s fitness. Similarly, increasing breakage rates also increases a crystal’s fitness (unless breakage occurs more often than growth).

In previous examples of zig-zag ribbon evolution [SW05b], tile sets (or “chemistry”) needed to be made more complicated in order to achieve more complex selection. In contrast, in biology a single chemistry has led to the evolution of more and more complex organisms. Such evolution has occurred because the fitness of a biological organism is a function of whether its genome contains a program that efficiently directs resource acquisition, development or relations to other organisms in its environment, and thus changes in the environment can drive the evolution of complexity (and *visa versa*).

Is there an analog of this mechanism for crystal evolution? That is, is there a single tile set that allows zig-zag crystals to achieve selective advantage in many different environments by performing functions adapted to their environment? While a zig-zag ribbon cannot direct any function except its own assembly, the fact that the assembly of tile crystals can perform universal computation [Win96] suggests that the answer could be “yes”.

An important element of the survival of a self-replicating system in the physical world is the ability to handle variation in the availability of raw materials needed for growth. While modern cells use genetic networks and signal transduction in order to respond optimally to available raw materials, here we describe how zig-zag crystals could evolve a simple “metabolism” by using tile assembly to compute which tiles to use for growth. This mechanism, consisting of a set of tile types and their environment, is too complex to be a model of real crystal growth processes. It is instead intended as a conceptual demonstration that it is possible for zig-zag crystals to evolve complex phenotypes.

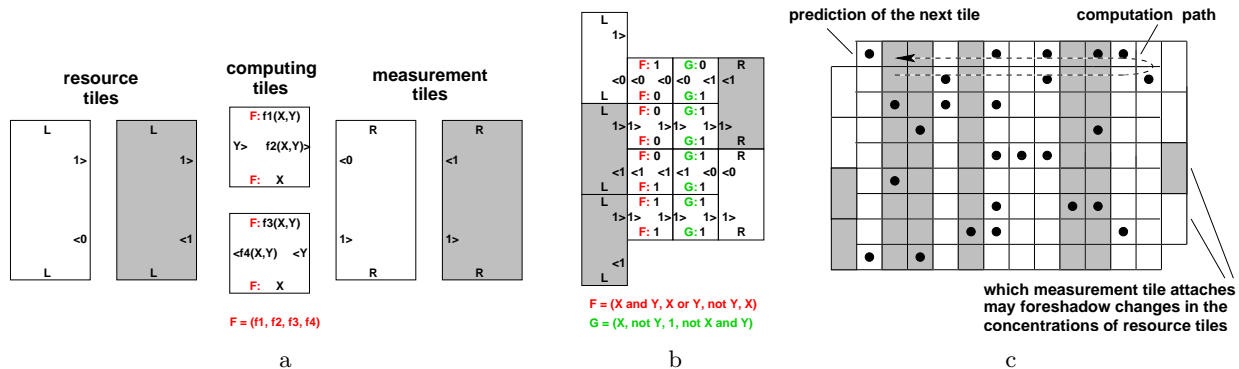


Figure 6.3: **Zig-zag crystals which exhibit metabolic control.** (a) A tile set for the evolution of metabolic control contains computation, resource, and measurement tiles. Computation tiles (middle) contain two kinds of labels: those on the left and right encode a boolean value, either 0 or 1, (shown as X, Y, or $f(X,Y)$) and the direction of assembly, either left or right. The labels on the top and bottom encode a gate type (**F** is shown) and a boolean input and output. Each gate type consists of four boolean functions : those passed to the top on leftward and rightward growth, and those passed to the left and right on leftward and rightward growth respectively. A computation tile has the same gate type on the top and bottom, thus ensuring that the sequence of gate types is copied from row to row. We assume that gate types corresponding to all combinations of four boolean functions are provided: there are 8×16^4 such tiles, which is admittedly quite large, although similar behavior should be possible with many fewer tile types. Resource tiles are boundary tiles to the left that have the same output value, a 1, but different input values. A crystal must provide binding sites for common resource tiles in order to grow quickly. Measurement tiles have the same input, but different outputs. These outputs can provide growing tile assemblies with information about the environment. “Smart” crystals use the information provided by the measurement tiles to predict which kind of resource tile is most available. (b) An example computation using the gates shown in (a). For each gate type, the four computations in parenthesis are the four computations for each gate type, in the order shown on the computation tiles in (a). (c) A large assembly consisting of the tiles in (a). The shading of computation tiles represents their gate type, which is passed from row to row and determines the genotype of the ribbon. Squares with dots represent tiles that output 0s, and squares with no dots represent tiles that output 1s. The 0s and 1’s represent the state of a computation, which is updated (but not necessarily copied) from row to row.

We consider a situation where tile resources may be limited and where the addition of a tile may provide information about the environment. In the examples presented here, boundary tiles are present at varying concentrations while the concentrations of non-boundary tiles do not vary.

Two types of boundary tiles, called measurement tiles (Figure 6.3a), initiate new rows from the right during upward growth. Both measurement tiles have the same input edge, but different measurement tiles have different output edges. Because measurement tiles share the same input edges, both available measurement tiles can attach to the right edge of the crystal. The chance that a particular measurement tile will attach is dependent on its concentration in solution. As will be described below, which measurement tile attaches determines the output edge that guides proceeding assembly of the crystal. We will assume that while the relative concentrations of measurement tiles change, their total concentration stays the same.

Another set of tiles may also become more or less available as time goes on. These are the resource tiles (Figure 6.3a). Changes in the concentrations of resource tiles may be correlated with changes in the concentrations of measurement tiles. While both measurement tiles have the same input edge, each resource tile has a different input edge. Only the resource tile that matches the left label on the binding site where a resource tile may be added can fit.

Figure 6.3a shows a set of “computation tiles” that perform boolean functions. In Section 6.2, we described how the requirement that a tile match the perimeter of an assembly by two edges in order to attach allows tile assembly to copy a sequence. A similar principle allows a crystal to perform local computations that modify the sequence: the two edges by which the tile attaches to the assembly serve as inputs, and the two edges of the tile that remain unattached serve as outputs [Win96]. These outputs then serve as inputs for future computation steps. On the computation tiles shown here, the label on the bottom of each tile encodes the gate type and an input value, and the label on the top encodes the same gate type as well as an output value. The left and right edges of the tile encode input and output values, depending on the direction of assembly. Thus, with each zig and each zag of growth, the sequence of gate types is copied verbatim from row to row, while simultaneously the sequence of boolean values are being processed according to the logic specified by the gates (Figure 6.3b). That is, the computations performed by tile attachment modify the sequence of boolean values from layer to layer, but the sequence of gate types is inherited intact. The concentration of the computation tiles remains constant over time.

The label on the output edge of an attaching measurement tile, either 0 or 1, will serve as an input for the attachment of the next tile, a computation tile. Conversely, the label that will be an input edge for a resource tile, either 0 or 1, will be an output of this series of computations. Thus, the order in which measurement tiles arrive at the right side serves as input to the computation tiles, which in turn produce outputs consisting of a series of input edges for resource tiles on the left side.

What kind of assembly is selected for in this environment? Fit assemblies are those which have the highest replication rate. If it is assumed that all assemblies split at the same rate, a fit assembly is one that grows quickly. Because the environment changes over time, the fittest assembly is the one that has a large *average* growth rate. The rate at which a new row is added, k_s , is the inverse of the total time that is needed to complete a row. The time needed to add each measurement and computation tile stays constant, because the rate of tile addition is dependent on tile concentration in solution, and these tiles are present in constant supply. (While the concentration of the two kinds of measurement tiles vary, their total concentration remains constant and both types will bind at a given binding site.) The time needed to bind a resource tile changes, however, and is dependent on the current concentration of the resource tile. Two observations can be made. First, a row is added more quickly if fewer computation tiles are used, and second, a row is added more quickly when the resource tile that is needed to complete the row is abundant. To achieve the first requirement, an assembly should be as thin as possible. To achieve the second, an assembly should correctly predict the abundant resource tile and assembly should produce a binding site for it rather than for a less abundant resource tile.

How can a program predict the future concentrations of tiles? The assembly of tiles transforms an input signal, a series of bits received from the output edges of the measurement tiles, to an output signal, a series of binding sites for resource tiles. This transformation depends on the sequence of gate types that is propagated in the crystal. A fit assembly produces an output signal that is the

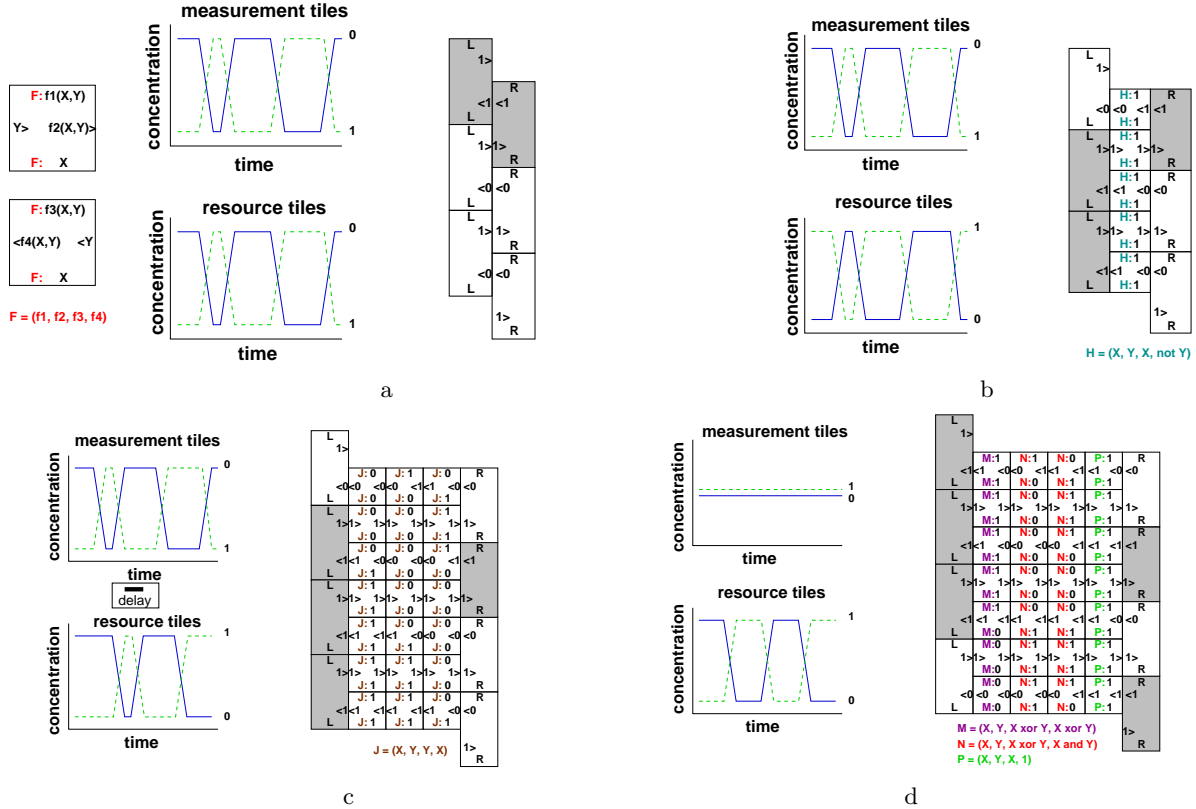


Figure 6.4: **Time varying tile concentrations and adapted assemblies.** Assemblies that grow from a set of tiles that encode boolean functions. Shown here are four environments in which the concentrations of tiles change over time, and an assembly that is well adapted to each environment. Selection favors crystals that correctly predict the concentration of sometimes scarce resource tiles. **(a)** Over time, measurement tiles and resources tiles have exactly the same time-varying concentrations. A well-adapted assembly asks for a resource tile with the same label as the measurement tile that was received. **(b)** Opposite resource and measurement tiles have the same concentrations. A well-adapted assembly uses a single computation tile that inverts the input from the measurement tile, so that the opposite resource tile can bind. **(c)** The kind of measurement tiles that are abundant is perfectly correlated with the type of resource tiles that will be abundant after a fixed time delay. A well-adapted assembly stores information about previous tiles and passes it across the assembly. If the time necessary to pass the type of measurement tile across the assembly is approximately the same as the delay in the correlation, the assembly will ask for the abundant resource tiles. A wider assembly produces a longer delay. **(d)** Measurement tiles provide no information, but resource tiles change regularly in a way that an assembly can predict. The assembly shown here uses a small counter [CRW04] to change its resource tile request every 8 layers. The left-most row remembers the current choice until the counter sends the message to change it. The rightmost column disregards the information provided by the measurement tiles.

same as the binding site of the more abundant resource tile¹. For example, when the input and

¹In growth that proceeds downward, rather than upward, the tiles labelled resource tiles function as measurement

output signal are time varying but identical, as in Figure 6.4a, an output signal that is the same as the input signal accomplishes this goal. When they are exactly opposite, as in Figure 6.4b, inverting the input signal, as is shown, accomplishes this goal. Thus, the most fit genomes in these environments are the empty sequence and the inverting gate respectively.

In some environments, an assembly that successfully predicts the identity of the more abundant resource tile must perform a less trivial computation. Figures 6.4c and 6.4d show examples, described below, of assemblies that do such computations. While a thinner assembly may add the computation tiles in its row more quickly, it would often ask for less abundant resource tiles. Thus, it would spend time waiting to bind these tiles, and therefore might grow more slowly overall than an assembly with more computation tiles that used abundant resource tiles.

A tile program that is well adapted to the environment in Figure 6.4c, where the output signal is a time-delayed copy of the input signal, consists of a series of “delay” gates. A delay gate passes the bit it receives from the previous row to the left and passes the bit received from the right to the next row. With a program consisting of n delay units, the crystal will respectively bind a resource tile with a 1 or 0 binding site $2n$ rows after receiving a corresponding measurement tile with a 1 or 0 output site. For a longer or shorter delay, more or fewer delay gates may be used.

When measurement tiles provide no information about the concentrations of resource tiles, as in Figure 6.4d, a successful assembly is one that ignores the information received from the measurement tiles and computes a set of outputs in a temporal pattern very similar to the changes in the abundant resource tile. The assembly shown uses its rightmost gate to discard the input from the measurement tiles. It uses a binary counter program [CRW04] (the center two computation tiles) to count to four over and over again. A counter consists of a series of exclusive or/and gates. The inputs to the counter come from rightmost digit of the counter (which is a 1 at each iteration) and the bottom edge of the counter tiles. At each iteration, the counter has two outputs – an output to the left which signals to the output tile to the left, and a set of outputs that become inputs to the next iteration of the counter. The “counter” receives its name because these outputs are, read as a binary number, one larger than the inputs. When the counter reaches its maximum value, in this case three, a one is output to the left. The left-most gate uses this periodic trigger to change its requested resource tile. The assembly shown asks for four 1-type resource tiles, then asks for four 0-type resource tiles, and repeats this cycle. If the requests are in tandem with the periodic change of resource tile type, the assembly will spend little time waiting for resource tiles, and will grow quickly.

The assemblies containing these programs will spend little time waiting for resource tiles; thus, they will grow faster than other assemblies of the same size that must wait for the right resource tile to become available. But will they grow faster than thinner assemblies, which don’t have to spend time adding computation tiles, even if they sometimes wait for resource tiles? In the examples provided, the answer is yes, if the total concentration of resource tiles is sufficiently low.

As an illustration, we compare the growth rate of the matched delay assemblies of Figure 6.4c with the growth rate of the “null assembly”, shown in Figure 6.4a, both growing in the delay environment of Figure 6.4c. Under what circumstances would a delay assembly of width k grow

tiles, and vice versa. Thus, both assemblies that can predict the resource tile types that are available from the measurement tiles, and those that can predict the measurement tile types that will be available from the current concentration of resource tiles will be selected for. Note, however, that gate types may be non-deterministic during downward growth, which could result in crystal growth stalling when a tile is incorporated that creates a binding site that matches no gate tile’s outputs. Therefore, consideration of downward growth rates is necessary for a complete evaluation of a crystal’s fitness; but we neglect it here to simplify the presentation.

faster than the null assembly? If tiles are added at an average of 1 per second, and resource tiles overall are D times less common than computation and measurement tiles, the growth rate of a perfectly synchronized delay assembly of width k “J” gates is $2k + D + 1$ per zig-zag. (A width k assembly is suited to a delay of $\Delta t = k(2k + D + 1)$ seconds.) Assuming that the type of resource tile that is abundant switches on average every s seconds, with $\Delta t \gg s$, the null assembly adds a zig-zag every $2(D + 1)$ seconds on average². When $2k < D + 1$ (and Δt is such that the assembly provides the right delay), the delay assembly grows faster. Clearly, this is true for large enough D .

Similarly, for the binary counter assembly (Figure 6.4d), the growth rate of the perfectly synchronized counter of width k is $2k + D + 1$ seconds per zig-zag³. The growth rate of the null assembly is $D + 1$ seconds per zig-zag—half the time for an average growth rate of $2D + 2$ seconds per zig-zag. Thus, when $k + D + 1 < 2D + 2$, the counter assembly is more fit. These examples illustrate that when D is large, that is, when the time spent adding computation and measurement tiles is negligible compared with the time spent waiting for resource tiles, performing the right computations can make an assembly more fit.

In order for a crystal that is good at predicting the current availability of resource tiles to be fit, it must be able to pass along its fitness to its descendant crystals. In this section, we have shown how every row of a crystal could contain the information for a simple program. Thus, when a crystal splits, each descendant contains the same program, which will also be executed. However, while the program is preserved by the descendants, the state of the program is not. The descendant crystals may start by executing the program from many steps ago. In some such cases, such as where a crystal is running a program to keep track of delays, the descendant crystal would then not grow well until it resynchronizes with the environment.

It might seem easier for an assembly to simply accept both kinds of prediction tiles. However, the chemistry of the tile set forbids this—1 and 0 labels do not match, and therefore cannot bind to each other. The fitness landscape is such that increases in fitness can only be achieved by good predictions. Thus, our examples address the question we set out to consider: how a fixed chemistry (a tile set) induces a selection pressure in which crystals must compute in order to be fit.

6.5 Evolution of Universal Sensing and Response

In the last section, we showed how a set of tiles that allow crystals to execute and replicate a program could, in the right environment, select for crystals performing a useful computation. However, the tiles shown in Figure 6.3a cannot simulate a Turing machine, and therefore cannot perform universal computation [Sip97]. Thus there are computations the tiles cannot express; such computations might be needed for accurate prediction in some environments. In this section, following [RW00], we describe a tile set that can simulate a Turing machine and how we can alter this tile set to perform the sensing of measurement tiles and binding of resource tiles described in Section 6.4. Because the Turing machine can compute any desired function, crystals grown using this tile set are capable of universal sensing and response. That is, in response to arbitrarily complex relationships between measurement and resource tiles, the fittest programs can also become arbitrarily complex.

²This estimate assumes that since $\Delta t \gg s$, at any particular time the resource and measurement tile concentrations are uncorrelated. So half the time, the resource and measurement tiles are compatible, and the crystal grows at a rate of $D + 1$ seconds per zig-zag; and half the time the resource and measurement tiles are not compatible, and the crystal essentially doesn’t grow.

³Note that counters whose natural period is slightly less than the period of the environment will easily remain synchronized with the environment, even when there is variation in the arrival times of the tiles being added.

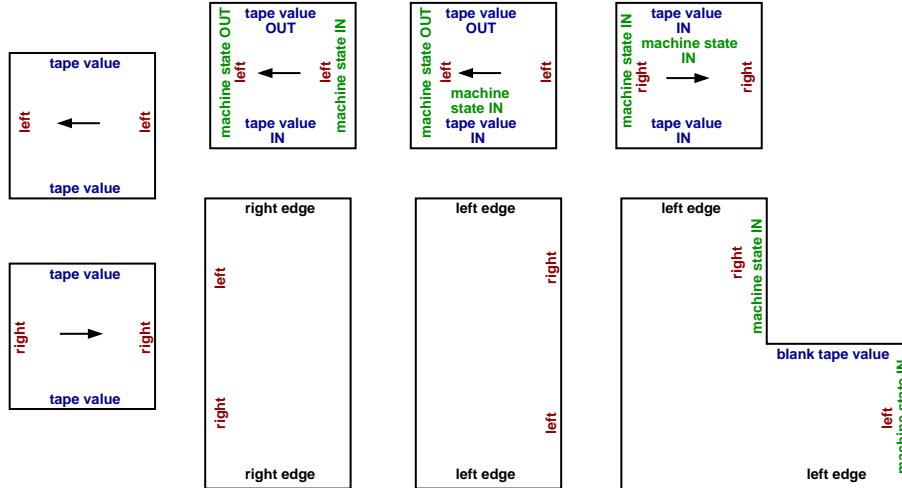
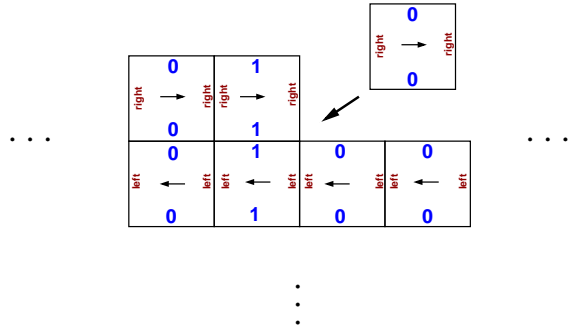


Figure 6.5: **Tiles for the simulation of a Turing machine.** A schematic for a set of tiles whose assembly can simulate the computation of a particular Turing machine on a tape consisting of a row of tiles. A tile type exists to copy each possible work tape value for the cases when assembly is proceeding both to the left and to the right. For each head state and input combination, tile types are needed for encoding the cases where the head is continuing to move in the same direction as the last step, is switching directions on the next step, or has switched directions on the last step. For head state and work tape combinations where the head should move to the right after computation, the directions of assembly will be the reverse of those on the tiles shown here. Left and right double tiles form the boundary of the work tape. When more work tape is needed to the left, an L-shaped tile can provide one extra work tape location on the next row (converse tiles exist to extend the work tape to the right). For a Turing machine that has k work tape symbols and s head states, $3ks + 2k + 2s + 2$ tiles are needed to simulate it. (A slightly more complex tile set can also shrink the width of the zig-zag when less work tape is needed.)

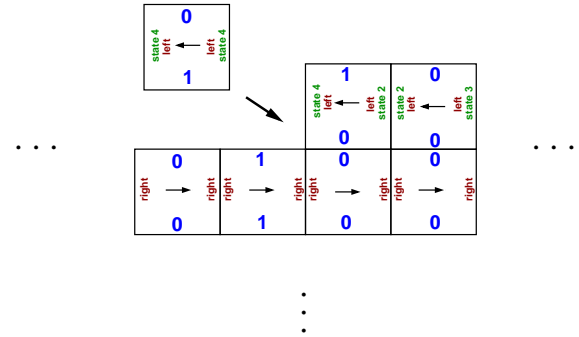
A Turing machine consists of a long work tape which has a sequence of symbols written on it, and a head that can be in a finite number of states. The head examines the work tape, one symbol at a time, executing a series of movements and updates to the symbols written on the work tape based on what it observes locally. At each state of the computation, the head is in a particular state and at a particular position on the work tape. The state and the symbol written on the work tape determine a symbol the head will replace the current work tape symbol with, the direction the head will move in (either one step to the left or right) and the next state for the head to enter.

The set of tiles that allows zig-zag ribbon assembly to simulate the execution of a particular Turing machine⁴ are shown in Figure 6.5. Each row of an assembly of these tiles represents the work tape at a particular time point in the computation. The design of the tile set is such that as assembly proceeds up the ribbon, most of the tape is copied from row to row (Figure 6.6a), but some cells are altered as directed by the Turing machine. For compactness and efficiency, all tape manipulations performed during a single unidirectional run of the Turing machine head occur in a

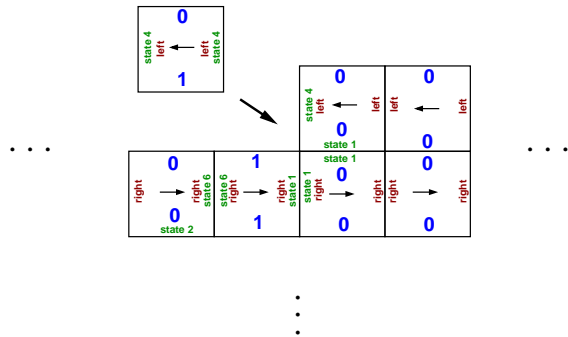
⁴While zig-zag assemblies copy layers on both growth fronts, for the purposes of our argument, we will assume that computation on zig-zag assemblies proceeds in only one direction (upward). While this is not necessarily so for the tile sets we describe, growth can be restricted to one direction by using a transformed tile set [Win06] that uses extra tiles to prevent growth in the wrong direction.



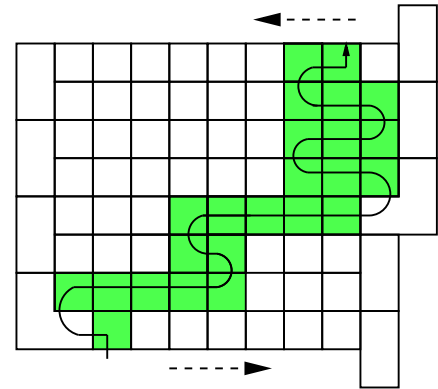
a At a position where the head is not present, an attaching tile copies the value on the work tape.



b When the head is an input from the right, computation proceeds because the attaching tile matches in the input state of the head. The output edge of this tile contains the next head state. If computation continues to the left, this head state is output to the left. The value on the work tape is updated according to the new value specified by the old work tape value and head state.



c If a head state and work tape value direct the head to move in the opposite direction as assembly is proceeding, the head state is output up, and computation occurs on the assembly of the next level, by attaching a tile that contains the head state on its bottom edge. Computation can then continue in the opposite direction (or switch directions again).



d An example of the progress of the head (shaded) of a Turing machine on a zig-zag that is simulating a Turing machine. White cells simply copy work tape values from the row beneath them.

Figure 6.6: Computation with the Turing machine tile set

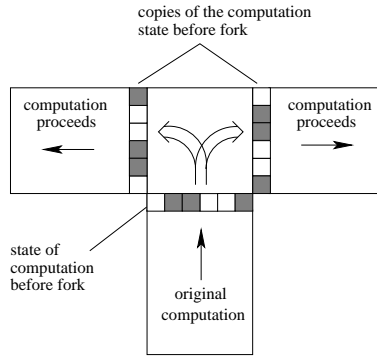


Figure 6.7: **Reproduction by budding.** While splitting is simple, assemblies might also program their reproduction time by “budding” instead of breaking. Budding assemblies may reproduce by first assembling a structure that allows computation to proceed in two directions.

single layer of the crystal. Thus, there is one layer per reversal of the Turing machine head.

At each step of assembly, exactly one location on the top row of the assembly has a place for a tile to join by two edges, and thus is available for assembly. At each such location, one of three things happens. If the position where the tile is to be added is not the position of the head, the attachment of a new tile will copy the tape from the last row. When assembly is proceeding to the right, the left edge will present a “right” label, and a tile can attach to a location along the top row if it directs assembly to the right and matches the tape value presented by the cell directly beneath it. Likewise, a tile can attach as assembly is proceeding to the left if it matches the “left” label presented by the right edge and matches the tape value presented by the tape (Figure 6.6a). These steps copy the values on the work tape.

At locations where the head is present, values on the work tape can change in subsequent layers. Here, a matching tile type not only matches the old work tape value and the direction of assembly, but also the current head state. The output edges of such a matching tile direct the new head state and the new value of the work tape at this location (Figure 6.6b). If the tape value and head state direct the head to move in the direction of assembly, the output edge in the direction of assembly (either to the left or right) encodes the new head state. If the tape value and head state direct the head to move in the direction opposite to the direction of assembly, the upward-facing edge of the attaching tile encodes the new head state (Figure 6.6c). Computation continues when assembly finishes in the current direction and zig-zags back to the column where the tile attached (Figure 6.6d).

Because a zig-zag tile set exists for the simulation of an arbitrary Turing machine, such a tile set exists for the simulation of a universal Turing machine. A universal Turing machine is a Turing machine that can simulate the execution of any other Turing machine when both the desired Turing machine and this Turing machine’s initial work tape are encoded on the universal Turing machine’s work tape [Sip97]. Small universal Turing machines exist that compute efficiently [Wat61] and simulate a Turing machine by encoding it on one part of the tape and a work tape on another, as illustrated in Figure 6.8a.

It is not hard to imagine using the measurement and resource tiles described in Section 6.4 instead of the single kind of boundary tiles shown in Figure 6.5 with the computation tiles that can simulate a universal Turing machine. It would also be possible to add a few tiles to input

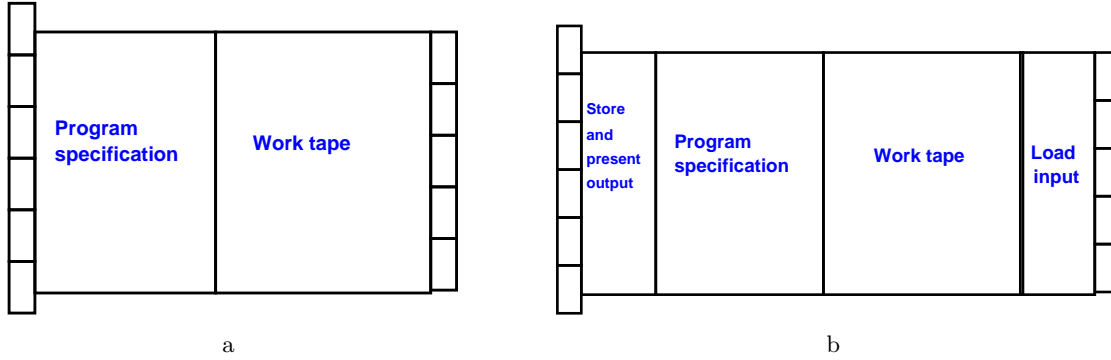


Figure 6.8: **Using a tile set capable of universal computation for metabolic control.** (a) The construction shown in Figure 6.5 shows how to translate an arbitrary Turing machine into a tile set that simulates it. Applying this construction to a universal Turing machine (UTM) yields a tile set that forms zig-zag ribbons of the type shown here: the program being executed by the UTM is stored in the ribbon separate from the work tape used by the program. (b) A tile set that simulates a universal Turing machine can be combined with boundary tiles that are either resource or measurement tiles. Tiles can also be added to load the input from measurement tiles onto the work tape and to output the desired binding site for the resource tile. With such a tile set, crystals that simulate a Turing machine that correctly predicts the concentrations of resource tiles over time would be most fit.

the measurement value and pass it onto the work tape, and likewise, to pass an output from the Turing machine to the left edge of the assembly as the binding type to present in order to bind a resource tile. The result would be assemblies with the structure shown in Figure 6.8b. With such a tile set, it could be possible for assemblies that computed and responded to any desired correlation between measurement and resource tiles to grow. Given an environment in which there is a complex correlation between the measurement and resource tiles that can be computed by a Turing machine, under conditions where the resource tiles are sufficiently rare, the fittest assembly would be one that exactly computed this correlation and asked for the appropriate resource tiles.

If the correlation between measurement and resource tiles were very complex, it is possible that by the time the assembly that exactly simulated the correlation finished computing it, the concentrations of the resource tiles would already have changed. In such an environment, a complex program for prediction may not be selected for. However, one can imagine an almost identical environment where that program would be selected for, in which the time the measurement tiles were present, the wait between presenting the relevant resource tiles, and the time the resource tiles were present were all longer by a constant factor. In an environment where the pattern of increasing and decreasing concentrations of measurement and resource tiles was sufficiently slowed down, the complex assembly would eventually be able to predict the right resource tiles in time, and would therefore be fit.

Since Turing-computable correlations have arbitrarily long history dependence, it may be difficult or impossible for a descendant crystal whose growth edge is at a previous state in the computation to resynchronize after crystal reproduction by fragmentation. A simple augmentation to the tile set rectifies this problem: a special head state triggers a “budding” process that duplicates and forks the computation (Figure 6.7). This produces two growth fronts that have the exact same

computation state as well as program.

A generalization of this type of algorithmically controlled morphogenesis was previously described in detail for investigating the Kolmogorov complexity of self-assembled shapes [SW04]. Since some shapes that could be created using this tile set may have selective advantage (particularly shear-prone shapes, for example, may be good at reproducing), evolution using such a tile set would result in selection for assemblies that contain programs for certain shapes. Descendants of these crystals will compute the same shape (perhaps starting from a random step in the computation) so that descendant crystals will also have selective advantage⁵.

6.6 Discussion

While it is possible to increase the complexity of selective pressures by increasing the complexity of tile sets used for growth, we've suggested here that tile sets exist that allow crystals to encode and run programs that can predict arbitrarily complex changes in the environment. Evolution using such a tile set as a medium could produce very complex sequences that grow well because they are particularly good at predicting changes in growth conditions. This suggests that crystal growth chemistries logically can support open-ended evolution with arbitrarily complex fitness landscapes.

Our argument that some assemblies will be more fit relies only on their growth rates, which are determined by how effective the assemblies are at predicting the availability of resource tiles. However, as Section 6.3 illustrates, to be fit, assemblies must not only grow quickly, but also shear frequently. Not enough is known about shearing frequencies to make confident predictions, as these rates are surely dependent on the kind of forces that produce shearing in practice. However, it seems safe to assume that wider zig-zag assemblies would shear less frequently than thin ones in most cases. If this were the case, of two assemblies that predicted the availability of resource tiles equally well, the thinner assembly would be more fit. Such an effect could actually have the effect of favoring correct *and concise* programs for prediction of resource tile concentrations, i.e., it is a natural implementation of Occam's razor. This scenario is somewhat akin to Solomonoff inference [Sol64] and Levin search [Lev73] in that short programs that produce correct results are found by biased random search.

However, our arguments in this paper are limited: they neglect several important real life features of crystal growth. For example, while the model used in this paper prohibits any tile that matches fewer than two bonds from attaching, such attachments occur at a rate dependent on the physical conditions of assembly. If this error rate is too large, two things can happen. First, particularly large programs that take a long time to run may have difficulty accurately computing without mistakes, and there may be a preference for so-called robust programs (if such programs exist for the tile set used), which compute an answer correctly even with a couple of assembly errors. Second, it may be possible for assemblies that ask for the absent resource tile to eventually attach the available but non-matching resource tile. Also, we have also not considered the effects of backward growth on the fitness landscape of crystals. For some tile sets, backward growth is not deterministic and quickly leads to a configuration that cannot be continued except by making an error, in which case backward growth stalls and can be neglected. For other tile sets, backward growth is deterministic and produces the same class of patterns that can be produced by forward

⁵In cases where a complete shape is necessary for selective advantage, it is also possible to use a construction where crystals can grow forward as well as backward, so that the parent crystal can grow back the piece of the shape that was lost during splitting [Win06].

growth. Although many tile sets are intermediate between these extremes, and backward growth should be considered, we expect that there are many tile sets for which crystal evolution can lead to complex fitness landscapes. Yet another simplification we have made is to ignore stochastic effects during assembly and how they affect the time at which an assembly asks for a resource tile. Again, we do not expect such considerations to significantly change our general conclusions.

Despite these and other limitations to our study as presented here, we believe that the qualitative features of the mechanisms that we describe should be preserved in a more realistic model of crystal growth. While the tile sets we described here are too complex to implement experimentally at this time, it is not unreasonable to believe that much simpler tile sets could produce complex adaptations in response to resource limitations. Our initial investigations into whether the ideas here could be tested in experiments suggest that tile sets containing as few as ten tile types could in fact exhibit complex evolution in an environment with constant tile concentrations, as will be described elsewhere. Such a small tile set and environment would be amenable to laboratory experiments to test whether non-trivial crystal genotypes that efficiently use available resources could evolve. While selection for larger and larger genotypes require more and more accurate copying of information [Eig71] and selection based on smaller fitness differences, error-resistant tile sets [WB04, CG05, RSY05], which contain more tiles but copy more accurately could be used. In principle these tile sets can reduce error rates as much as is desired.

One might think that for the fittest species in a given environment to be arbitrarily complex, the environment must be equally complex. But it is not immediately clear whether this is actually the case. Since there is no fastest algorithm for some functions [Blu67], this would seem to imply that in a crystal growth environment that requires crystals to compute such a function, evolutionary pressure favoring speed would lead to ever-increasing complexity. On other hand, faster programs are in general larger. So while it is possible that evolutionary pressure might favor faster algorithms, the crystal must also be wider in order to contain the complex program and copying the program slows down the crystal's growth. This trade-off is also present in biological, *in vitro*, and artificial evolution.

It is also interesting to ask about the *minimal* requirements for open-ended evolution of crystals. Is it possible that there is a “universal” environment where tile concentrations are constant or vary in a simple periodic pattern, but open-ended evolution is still possible? To answer such a question, it may be worth considering effects that are beyond the simple model of crystal growth and replication kinetics used in this paper. We considered only fitness of species for growth in the exponential phase, where resources are not depleted by other crystals, and where there is no interaction between crystals. Relaxing these assumptions may allow other interesting routes to complexity. For example, could competition between crystals for tiles lead to an “arms-race” of more and more complex strategies for growth? Conversely, might symbiosis between crystal programs lead to interesting phenotypes not explored here? Might a crystal growth chemistry support parasitic crystals that evolve to grow by using existing crystals rather than single tiles? Supporting this last possibility, initial experiments with DNA crystals indicate that joining of crystals does occur [ENKF04].

Chapter 7

Evolution of Complex Crystal Sequences from Simple Parts

Abstract

In 1966, Graham Cairns-Smith proposed that the first genetic material on earth was not organic, but crystalline, and resided in the particular morphologies of clays. One of the primary objections to this theory is that it is not clear how non-trivial crystal genotypes could be selected for. It is therefore of interest to ask whether crystals and physical conditions for their assembly exist where complex crystal morphologies could evolve. DNA tile crystals are particularly suitable for this kind of investigation because DNA tile monomers are programmable—new molecules with particular affinities can be easily designed and synthesized. While it has previously been suggested that arbitrarily complex morphologies of DNA tile crystals could form as the result of an evolutionary process, the crystals which could develop these morphologies contain thousands of monomer types, and are therefore too complex to assemble or study. Additionally, the physical conditions under which this process might occur would be difficult or impossible to reproduce experimentally. Here, I suggest how evolution of crystals produced by a small set of tiles could yield crystals with complex patterns under achievable environmental conditions. The crystals consist of a set of 12 tiles that copy a 2-dimensional pattern along a ribbon-shaped crystal. Each of these tile sets can form some large, complex crystals which only rarely use specific molecules during growth. Simple, thin crystals constructed from these tiles must use the specific molecules more often. When these molecules are available only at very low concentrations, the large, complex crystals that only rarely use these molecules grow faster than thin, simple ones. Thus, large, complex crystals are selected for. Because the number of these molecules that are required for growth decreases as the width of the crystals increases, by making the concentration of these required molecules arbitrarily small it is in principle possible to select for arbitrarily large, complex crystals. I describe a family of tile sets whose members contain tile sets that meet the above criteria. I investigate the crystal evolution that occurs with these tile sets using both a simple analytical model of crystal growth and stochastic simulations. The molecules and the environments described are sufficiently simple that evolution of complex crystal forms could be tested in the laboratory in the near future. The described mechanism for evolution is also sufficiently general that it may apply to other DNA crystal forms such as nanotubes, or even to natural clay crystals, suggesting that the crystals Cairns-Smith imagined might have evolved complex forms because of a relative lack of some crystal monomer types.

7.1 Introduction

How life on earth originated remains a mystery for two primary reasons. First, I have no idea how a self-replicating chemistry capable of evolution came about spontaneously. Second, I do not understand what features of a simple self-replicator could allow Darwinian evolution to produce the complexity found on earth today.

The question of how self-replication and evolution first arose has been the subject of much previous work. The oldest proposal of the form of the first life is a self-replicating molecular aggregate [Opa03, SBEDL01]. While the self-replication of lipid vesicles has been investigated in the laboratory [WWF⁺94], there are relatively few ideas about how such aggregates could evolve without genetic material [CRS04]. Other scientists have proposed that life might have originated with a set of molecules whose cross-catalytic activity would produce an autocatalytic cycle [FKP86, Wäc88, Wäc90]. While mathematical analysis of models of these systems has suggested that they have the ability to evolve [Eig71, Kau93, SBEL00], experimental demonstrations have been restricted to either systems for which catalysis does not form a complete cycle [HW], or to systems that contain autocatalytic cycles that do not obviously appear capable of non-trivial evolution [vK86, LGMKs96, LE03]. The most well-accepted hypothesis about life that preceded the current DNA/RNA/protein life is the RNA world hypothesis, which proposes that genetic information was once stored by RNA molecules which could catalyze their own replication. While it seems plausible that RNA sequences capable of catalyzing their own replication [JUL⁺01, PJ02] exist, template-based RNA replication has never been conclusively demonstrated. It has also been suggested that clay crystals could replicate information [CS66] but this has not been demonstrated, although initial work with DNA crystals suggest that replication and evolution of very simple sequences could be feasible [SWb].

Because it is still far from clear what the chemical properties of the first life were, it is even more difficult to make judgements about how its evolution might have led to more complex life forms. This work is further complicated by the absence of self-replicating chemical systems with which to do experiments. Experiments on RNA molecules [LCMY03, CODS04] have begun to map the distribution of catalytic function over sequence space [Smi70]. The knowledge necessary to speculate about RNA evolution, however, is not yet available. Simulations of artificial chemistries [Fon92] and of computer programs [Ada98] have demonstrated evolution, but in general so-called “open-ended evolution” where evolution creates new complexity without limit is viewed as an open problem [BMP⁺00]. Thus, while many scientists assume that self-replication of combinatorial genome usually leads to open-ended evolution, limited work suggests that open-ended evolution is not a trivial consequence of these phenomena.

In this paper, I investigate whether the evolution of self-replicating DNA crystals could lead to the assembly of complex crystals and possibly to open-ended evolution. DNA crystal self-replication is a proposed method of enzyme-free, chemical self-replication and evolution inspired by Graham Cairns-Smith’s proposal that life originated in the self-replication of clay crystal morphologies [CS66]. In DNA crystal self-replication, clay crystal monomers are replaced by DNA tiles [FS93], crystal monomers consisting of DNA strands, with binding sites for other monomers consisting of 4 short segments of single-stranded DNA.

The DNA crystals that I study here are ribbon-shaped. By design, they grow by adding tiles that copy the existing arrangement of tiles at each of two ends. Thus, each layer of the crystal contains the same piece of information. Sustained growth produces multiple copies of the sequence along the ribbon, but because the ribbon can only grow at each end, the number of copies the

ribbon contains does not change its growth rate. Splitting the crystal into pieces (as for example with mechanical force) produces new crystal growth fronts which accelerate the rate at which a sequence is copied.

Because replication of a crystal sequence requires both growth and splitting, crystals must both grow and split quickly in order to reproduce quickly. Previously, it was shown that when crystal breakage occurs more slowly than growth, the replication rate of a crystal sequence is the geometric mean of the rate at which the crystal’s genome is copied and the rate at which the crystals are split into two [SWa]. If periodic, violent force is used to split crystals in two it is reasonable to assume that crystals all break into pieces when this force is applied, and therefore that all crystals break at approximately the same rate. Thus, in the paper, I assume that a crystal which is able to copy its sequence fastest also replicates the fastest, and therefore is selected for.

To determine which crystals copy their sequences the fastest I use a previously described tile assembly model [RW00], a generalized crystal growth model that has been used to study algorithmic self-assembly of synthetic DNA tiles [WLWS98, RPW04, BRW05]. In this model, crystal monomers are considered to be square or rectangular tiles with each unit edge labeled to indicate whether it attaches to other monomers. Crystal growth proceeds by accretion, with single tiles being added to the crystal at sites where they make a sufficient number of contacts. In this paper I consider a version of this model in which (a) a tile may be added to a site if labels on at least two edges match those presented by the crystal at that site, and (b) monomer tiles arrive at potential binding sites with a frequency proportional to their concentration in solution. Occasional violations of rule (a) are referred to as “mutations”. A system consists of a set of tiles along with the concentrations of those tiles in solution. Because of the particular choice of matching rules, each tile set implicitly determines what arrangements of tiles can grow as crystals.

I investigate the behavior of tiles that form crystals which copy a 2-D pattern of tiles of fixed width. Each fixed-width column in the pattern can be viewed as a step in the history of a particular model of computation, the cellular automaton [FPU55, vNAWB66]. The cellular automaton is a simple model of computation in which a tape consisting of discrete cells is updated in parallel based on a series of local rules. While in many instances the tape is infinite, here the tape is limited by the width of the crystal. The attachment of a tile can be viewed as advancing the state of the computation at one cell in the cellular automaton’s tape. The identities of the edges to which the new tile attaches are the “inputs” to the local rule function which advances the computation, and the identity of the tile which matches these edges are the “outputs”¹. The assembly of an entire column advances the computation by one step. Repeated addition of columns will eventually result in the repeat of columns, which will continue as each state in a repeating cycle of states is visited in order. For each kind of computation rule and width of the crystal, there may be several such cycles [WL92]. Thus, for a particular tile set which determines the computation rule, the width of the crystal and the particular cycle determine the information copied by the crystal.

In the ribbon tile sets that I study, there are three kinds of tiles—“rule” tiles, which may be used in the middle rows of the ribbon, “top” tiles and “bottom” tiles, which fit only on the top and bottom rows of the crystal respectively. There are two kinds of top tiles and two kinds of bottom tiles. Each pattern that can be copied uses a certain percentage of each of the two kinds of top tiles and likewise, a certain percentage of each of the two kinds of bottom tiles. Under conditions where one type of top or bottom tile is present only at very low concentration (while the other is present

¹While zig-zag crystals can elongate in both directions, the computation described here may be irreversible. While tile sets exist that prevent backward assembly [Win06], in many cases it stalls.

at a normal concentration), the time it takes a crystal to grow is dominated by the time it takes to add this one type of tile. Thus, crystal patterns that use fewer of this tile on average grow faster and are evolutionarily fit. In many tile sets, I find that progressively wider, more complex crystal patterns use progressively fewer of one type of top or bottom tile per row. Thus, these wider, more complex crystals are selected for.

In this study, I define the complexity of a DNA ribbon crystal as the width of the crystal. Strictly, it would be more correct to describe the complexity of a crystal as the algorithmic (Kolmogorov) complexity [LV97] of describing one of its columns, since one column determines the current state of computation, and therefore all future states of computation. While the algorithmic complexity of a crystal does not grow monotonically with the width, the algorithmic complexity is bounded by, and often attains the value of the width of the crystal. In general, if crystals grow arbitrarily large, they also grow arbitrarily computationally complex [LV97].

I investigate tile sets for zig-zag ribbon crystals of DNA in this paper. These crystals have been extensively characterized [SW07, BSRW07] and limited self-replication of zig-zag ribbon crystal genotypes as been demonstrated [SWa]. Thus, it should be possible to relatively easily test the predictions made here about evolution of complex crystals in the laboratory.

While more investigation is needed, I expect that there are other kinds of synthetic crystals, including DNA tile nanotubes [RENP⁺04, MHM⁺04] for which a similar kind of evolution could occur, perhaps even more simply. Perhaps more speculatively, there is reason to believe that these same principles could apply to clay crystals, and therefore have induced the evolution of complex crystals in a pre-biotic environment, directly addressing criticisms of the crystal origin of life.

7.2 A Growth Model for Zig-Zag Crystals

The zig-zag tile sets I investigate produce crystals that process information along the length of the crystal. Each tile set includes a group of square tiles and a set of rectangular tiles called *double tiles*. Logical representations of an example zig-zag tile set are shown in Figure 7.1a. When growth occurs according to the tile assembly model [Win96], only tiles that match at least two edges can attach (Figure 7.1b.) There is therefore just one location where a tile that may be added at each end of the ribbon. The addition of one tile creates a new site for a tile to be added; tiles are added to a ribbon in a zig-zag pattern shown in Figure 7.1c.

The requirement that a tile attach by at least two bonds means that it must match both its vertical neighbor, either above and below, and its horizontal neighbor, to the left or right. While several tile types may match the vertical or horizontal neighbor, only the tile type that already appears in the row will match the horizontal neighbor (Figure 7.1b.) Thus, the addition of each new row to the assembly in Figure 7.1c can be viewed as the copying of the information in the previous row. Information is therefore inherited from layer to layer.

The same tile set can form many different kinds of crystals, each of which contains a sequence copied during assembly—the tile set shown in Figure 7.1a can copy binary sequences of arbitrary width (Figure 7.1d shows examples.) The same process of zig-zag crystal growth can be used to copy any of these sequences. Thus, one small set of tiles can copy any one of an infinite number of sequences. As an example of how the set of tiles change the alphabet of sequences that can be copied, with the addition of the two kinds of tiles shown in Figure 7.1e, sequences in a tertiary alphabet could be copied.

In crystal evolution, many different crystal sequences are copied in the same test tube by the

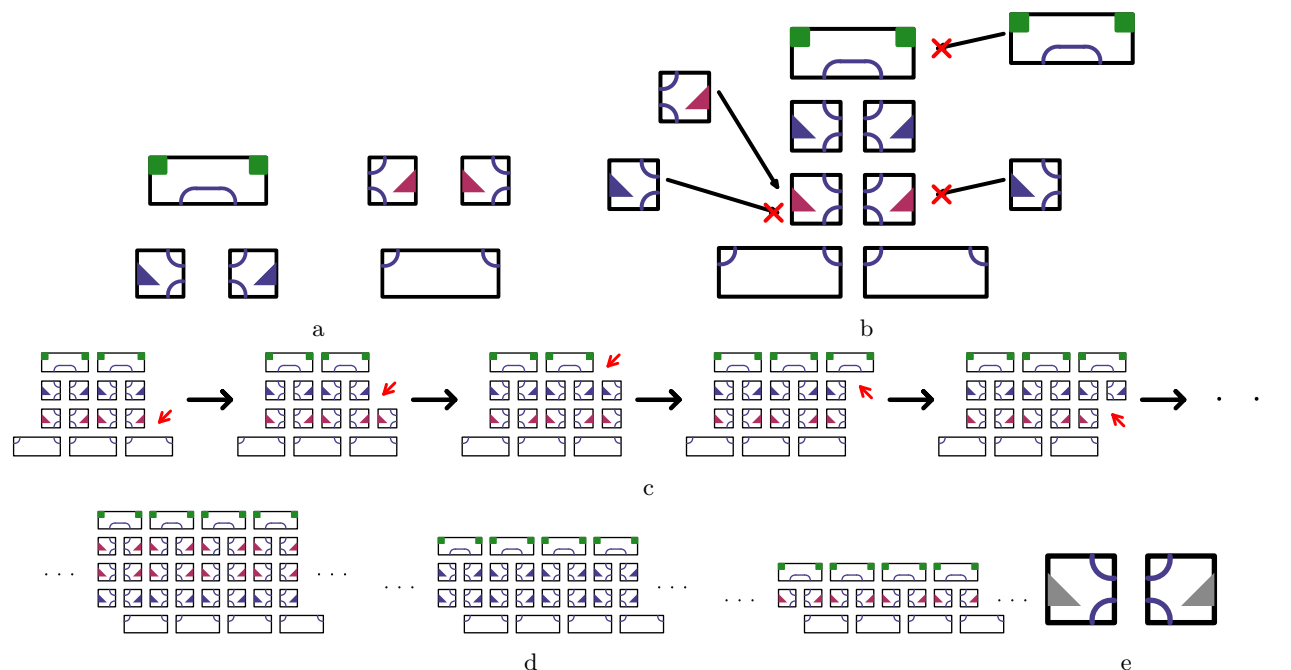


Figure 7.1: **The zig-zag tile set.** (a) A basic width-4 zig-zag tile set consists of six tile types. Tiles cannot be rotated. (b) The tiles in (a) can form the assembly shown here. Tiles can attach to the crystal only where they match at least two edges of the crystal. (c) The process of zig-zag crystal growth. At each step, a new tile may be added at the location designated by the small arrow. Two alternating tiles in each column enforce the placement of the double tiles on the top and bottom, ensuring that growth continues in a zig-zag pattern. While growth on the right end of the molecule is shown here, growth occurs simultaneously on both ends of the molecule. (d) Other assemblies that can be formed from the tile set in (a). These crystals can also grow by adding tiles in zig-zag order, as shown in (c). The tiles shown in (a) can form crystals carrying any binary sequence.

same set of tiles. As in any evolutionary process, faster growing (and better splitting) crystals are fitter and become much more frequent than slower growing crystals. Periodic errors in assembly will create new crystal types, and if crystals are periodically flushed from solution, less fit crystal morphologies will disappear. In the case of the tile set above, if tiles with pink triangles are present in higher concentration than tiles with blue triangles, sequences containing mostly or only pink tiles will be fitter and eventually dominate the solution. In the opposite case, if tiles with blue triangles were more highly concentrated in solution, sequences consisting mostly or only of blue tiles will be selected for. Thus both the tile set, which determines the alphabet of patterns that can be copied, and physical conditions such as tile concentration determine the results of an evolutionary process.

Figure 7.1a shows a tile set that copies a simple pattern; a repeating pattern of either blue or pink triangles in each row. It is also possible to build similar crystals that copy not a sequence but a repeating pattern of vertical sequences. Figure 7.2a shows such an example tile set. Like the tile set shown in Figure 7.1a, this tile set can copy an infinite number of patterns. But unlike the latter tile set, this new tile set copies patterns consisting of more than two columns. Figure 7.2b shows some example patterns that are repeated during growth. In a long crystal, copies of these patterns

appear adjacent to each other—the binding sites along the right edge exactly match the binding sites along the left edge.

In this paper I study the evolution of ribbons produced by tile sets like the one shown in Figure 7.2a when one or more of the rectangular boundary tiles is rare or unavailable in solution. In the next section I describe the set of tile sets I will consider in detail and their relationship to a well-studied model of computation, the one-dimensional cellular automaton [Wol02, Coo04].

7.3 Finite Cellular Automata

A cellular automaton is a model of computation that proceeds on a discrete, usually infinite, lattice. Unlike the Turing machine model of computation [Sip97], the results of computation do not determine which cell is updated next; instead all cells are updated in order (synchronously), or in some models, at random (asynchronously). The history of the computation of an n -dimensional cellular automaton can be recollected on an $n+1$ -dimensional lattice, in which the $n+1$ st dimension is time. Thus, the $n+1$ -dimensional lattice stores the state of the computation after each series of updates in the synchronous case and after each update in the asynchronous case. Here I consider a variant of cellular automaton I will call a **1-d zig-zag cellular automaton**.

Computation on a 1-d zig-zag cellular automaton takes place on a lattice of finite 1-d tape with two boundaries. By convention, the computation tape is a vertical column, and the horizontal axis of a 2-d lattice showing the history of a computation is time, with time proceeding to the right. Updating of cells occurs synchronously and proceeds from the bottom edge to the top edge at odd time steps and from the top edge to the bottom edge at even time steps. The input to each cell consists of two values from a fixed alphabet \mathcal{A} : a “right” value and either a “top” or “bottom” value. During updates proceeding from bottom to top, the new “top” value is a function $f_1 (\{\mathcal{A}, \mathcal{A}\} \rightarrow \mathcal{A})$ of the cell’s old “right” value and of the cell below the current cell’s new “top” value. The new “right” value is a different function, f_2 of these same values. Correspondingly, during the updates from top to bottom, two functions f_3 and f_4 that operate on the cell’s old “right” value and the cell above’s current “bottom” value determine the new “bottom” and “right” values respectively.

At the end of the update bottom to top, the last “top” value is used to determine the boundary “bottom” value that the top-most cell will use on the downward update. A fifth function, $f_5 (\mathcal{A} \rightarrow \mathcal{A})$ determines the new “bottom” value given the boundary “top” value of the previous row. Likewise, the last function, f_6 , determines the next “top” value from the last “bottom” value at the end of a series of updates from top to bottom. Figure 7.2c shows an abstract diagram of the tiles that can implement a zig-zag cellular automaton.

Because computation takes place on a finite lattice, there are only a finite number of states of the tape that are possible — 2^{n+2} on a tape of width n (updates can either be upward or downward, n cells have a right value and there is one boundary condition). Thus, a state must repeat eventually. Once a state is repeated, it must be repeated an infinite number of times. That is, there is a “cycle” of states that are entered over and over as computation proceeds. If computation is irreversible, some states may be “transient,” and be on the path to a repeating cycle, but never themselves repeated (for lots of details about the dynamics of cellular automata and illustrations of their state spaces, including cycles and transient states, see [WL92]).

A zig-zag cellular automaton computation can be implemented by a set of 12 tiles: four rule tiles to encode the four possible binary inputs and their respective outputs in each of the up and

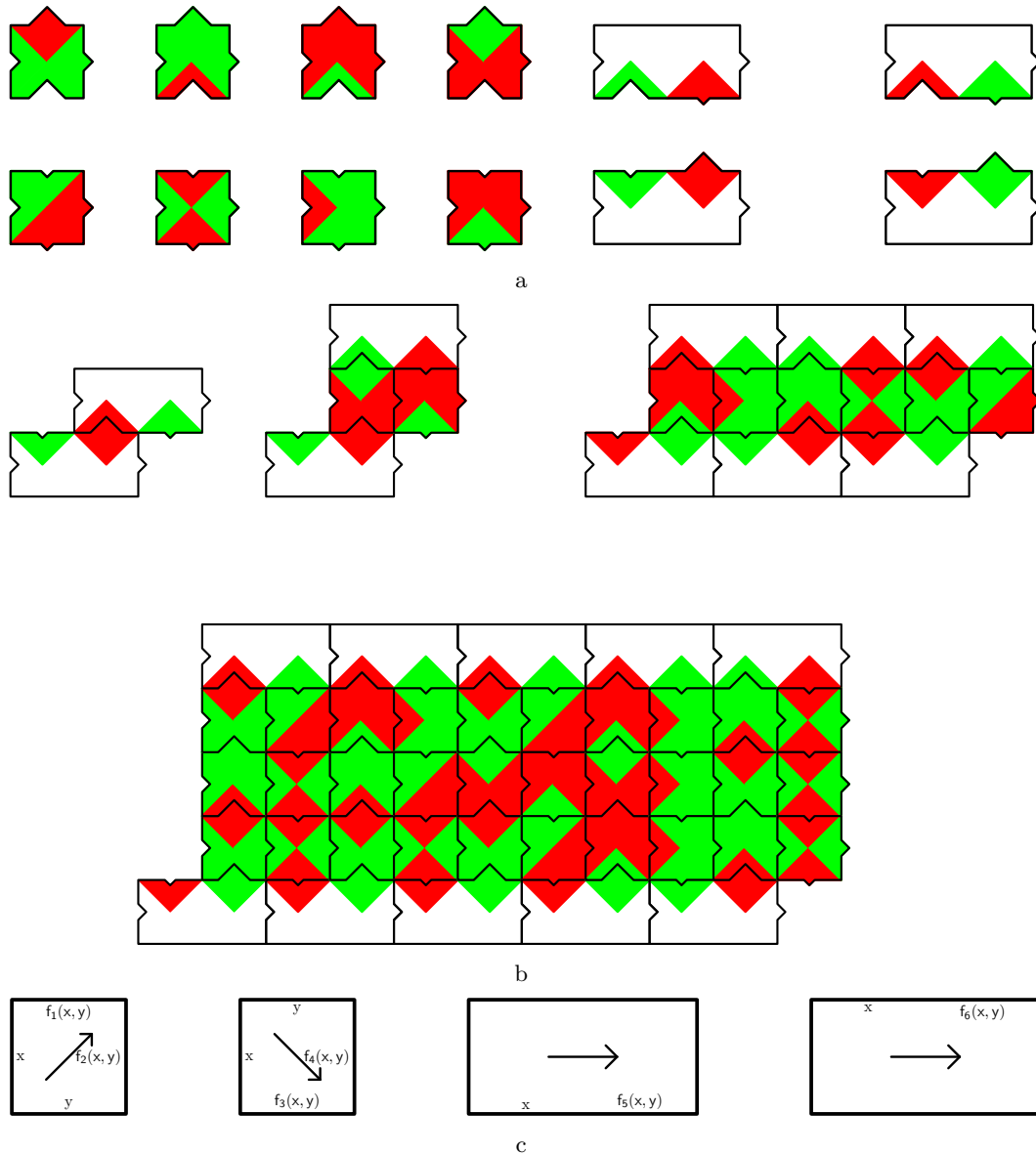


Figure 7.2: **Tiles to implement a cellular automaton.** (a) An example tile set that copies a 2-D pattern instead of a linear sequence along the ribbon. Green matches with green and red matches with red. (b) Example assemblies formed by the tile set shown in (a). Each piece shown contains exactly one iteration of a pattern that is repeated through continued growth of the crystals. (c) Abstract representation of tiles that implement a zig-zag cellular automaton. There are four tiles that can compute upwards and four tiles that compute down. There are two top and two bottom tiles. Six Boolean functions define the behavior of the tiles as they assemble.

down directions and four boundary tiles, consisting of two top tiles with the inputs 0 and 1 and their respective outputs, and two bottom tiles with 0 and 1 inputs and their respective outputs. Figure 7.2a contains the tiles to simulate one such automaton and Figure 7.2b show examples of computation which cover one cycle.

Tiles that implement a zig-zag cellular automaton assemble a repeating pattern, just like the tiles that copy a repeating sequence within a ribbon. The class of tiles that implement a zig-zag cellular automaton can be viewed as a generalization of the set of tile sets that simply copy a sequence at each row, where the latter are zig-zag cellular automata with many cycles of size 2 (cycles are of size 2 to enforce the staggered placement of the double tiles, ensuring zig-zag growth.) Such tile sets will be referred to henceforth as **copy tile sets**.

In contrast to the copy tile sets, a zig-zag automata tile set contains two kinds of each boundary tile—one with a 0 input and the other with a 1 input. In many cases, fast growth of a zig-zag automata tile set is still possible when one of these tile types is very rare because the tiles can form patterns that either do not use or only very rarely use the rare tile.

How does the percentage of rare tiles that are used affect the growth rate of a crystal? In our model, the time it takes a tile with concentration $[z]$ to attach is $\frac{1}{k_f[z]}$, where k_f is the forward rate constant of attachment, independent of tile type². Let's consider a situation where the concentration of the rule tiles is $[r]$, the concentration of the commonly available top and bottom edge tiles $[e]$ and the concentration of the rare edge tile $[q]$. The time it takes on average to copy each column of tiles in a cycle that contains w rule tiles per column, c common edge tiles and u uncommon edge tiles is

$$\frac{1}{k_f} \left(\frac{c}{2(c+u)[e]} + \frac{u}{2(c+u)[q]} + \frac{w}{[r]} \right) \quad (7.1)$$

when the concentration of both edge tiles is the same, $[q] = [e]$, and the time it takes to attach a row of tiles grows with the number of rule tiles in each column of the crystal. However, when $[q]$ is very small, the middle term dominates. In this case, a wider crystal with w_1 columns, c_1 commonly available edge tiles, and u_1 rare edge tiles can grow faster than a thinner crystal with w_2 , c_2 , and u_2 rule, common edge and rare edge tiles, respectively, if the following equation is satisfied:

$$2(w_1 - w_2) \frac{1}{[r]} > \left(\frac{c_2}{c_2 + u_2} - \frac{c_1}{c_1 + u_1} \right) \frac{1}{[e]} - \left(\frac{u_1}{c_1 + u_1} - \frac{u_2}{c_2 + u_2} \right) \frac{1}{[q]}. \quad (7.2)$$

Thus when $\frac{1}{[q]} \gg \frac{1}{[e]}$, a wider crystal can grow more quickly if it uses sufficiently fewer of the rare edge tiles. Further for any case where $\frac{u_1}{c_1 + u_1} < \frac{u_2}{c_2 + u_2}$ it is possible to find an environment (in terms of $[r]$, $[e]$, and $[q]$) where the wider crystal grows more quickly³

Thus, in some environments wider crystals which use proportionally less of one kind of the boundary tile than thinner crystals do are selected for. In the next section I search the set of

²In reality, attachment is reversible, and the time for attachment must incorporate a temperature-dependent detachment rate. An analysis of the time for a group of tiles to attach that takes into account the fact that tile attachment is reversible should change the quantitative results I derive here, but not the essential conclusion: A wider crystal can grow faster on average than a thinner one if the wider crystal's pattern contains fewer tiles that are available at a very low concentration.

³It is not hard to recreate such an environment in the laboratory. While it is in general difficult to get the concentrations exactly right in an experiment, it is possible to use concentrations for which a small amount of experimental error would not change the results because the selective advantage of the wider crystal grows with decreasing $[q]$.

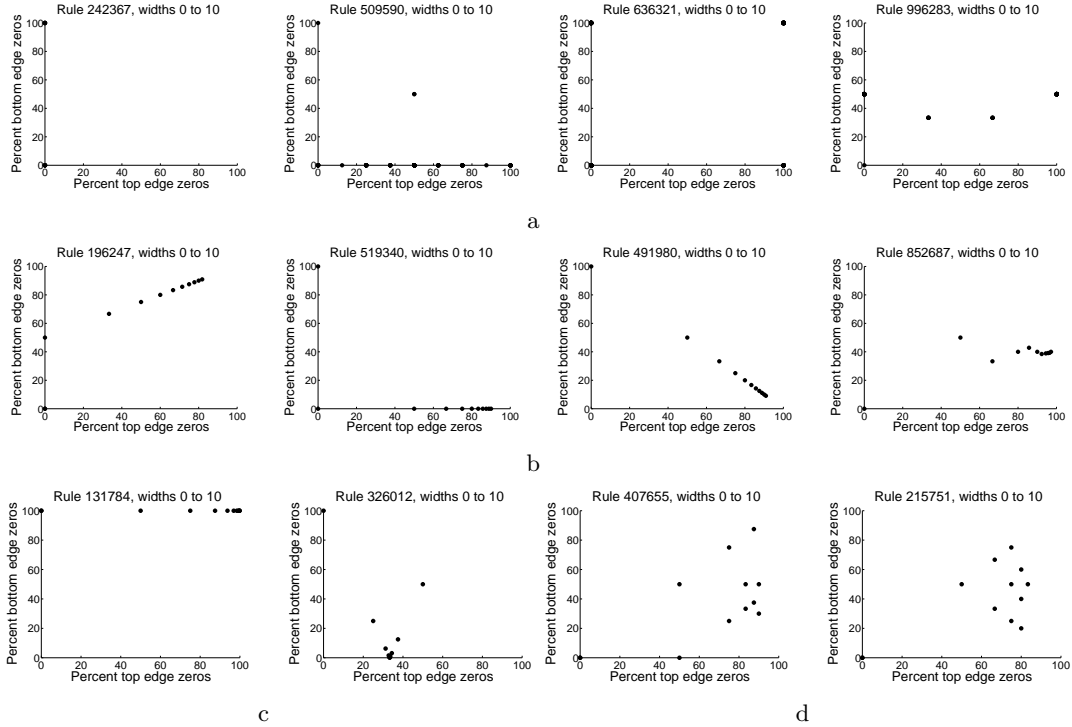


Figure 7.3: **Boundary tile usage by zig-zag cellular automata tile sets.** The rule number is an encoding of the six Boolean functions. **(a)** Usage of boundary tiles by the smallest cycles of 4 random rules. Many rules have consistent boundary tile usage for all widths simulated, but some have irregular patterns of usage. Most of these tile sets would not produce wider assemblies if one or more boundary tile types were restricted, but might evolve under other types of restrictions. **(b)** Assemblies for which growing wider reduces the usage of one or more boundary tile types by a linear ratio. **(c)** Assemblies for which growing wider reduces the usage of one or more boundary tiles exponentially. **(d)** Tile sets for which growing wider occasionally reduces the usage of bottom tile with a 1 input. It is likely there are many rules similar to these, but the screening of rules I did initially was restricted to tile sets that produced attractors that used new percentages of top and bottom boundary tiles with every larger width.

binary zig-zag cellular automata tile sets for tile sets that have crystals with this property. I find there are tile sets where the addition of each additional row of rule tiles allows the tiles to form a crystal which uses proportionally fewer of one or more kinds of boundary tiles than all thinner crystals.

7.4 Evolution of Zig-Zag Crystals Encoding Binary Cellular Automata

A zig-zag cellular automata tile set is defined by 4 two-input Boolean functions and 2 one-input Boolean functions (Figure 7.2c). There are 16 two-input Boolean functions and 4 one-input Boolean functions, so there are a total of $16^4 * 4^2 = 1048576$ such tile sets.

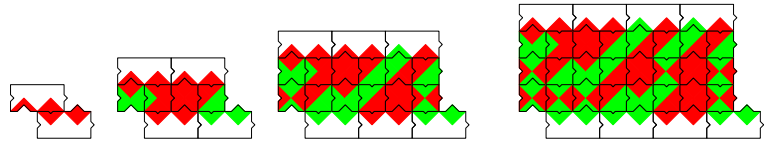
Since there are too many tile sets to look at exhaustively, I examined the usage of boundary tiles in crystals containing between 0 and 10 rows of rule tiles for 4 randomly chosen rules. The results are shown in Figure 7.3a. In the first example, all cycles use only tiles with a “1” input on the top and either all “0” or all “1” input tiles on the bottom. Restricting the concentration of one kind of bottom tile (say the “0” input tiles) would select for assemblies that use only “1” tiles. Because relatively thin crystals (less than 10 rows of rule tiles) use none of the “0” input tile, it would not be expected that crystals wider than this could be selected for by reducing the concentration of this tile. The concentration of edge tiles used by the crystals in the other 3 examples follow the same sort of pattern.

Tile sets for which successively wider crystals may be better adapted will, with successively wider crystals, have patterns of states that use a percentage of each kind of boundary tiles that no thinner crystal uses will appear. To identify some of these rules, I surveyed all rules and identified those for which new such patterns appear with each larger width for crystals containing between 0 and 5 rows of rule tiles (since the number of states grows exponentially with width, surveying all rules for wider widths was too computationally intensive.) This survey produced about 45,000 rules. A set of 100 rules randomly selected from this set produced a variety of candidate rules for which the restriction of some boundary tile types could induce evolution of wider crystals. The boundary tile usage of cycles produced by rules of these two types are shown in Figures 7.3(b-d).

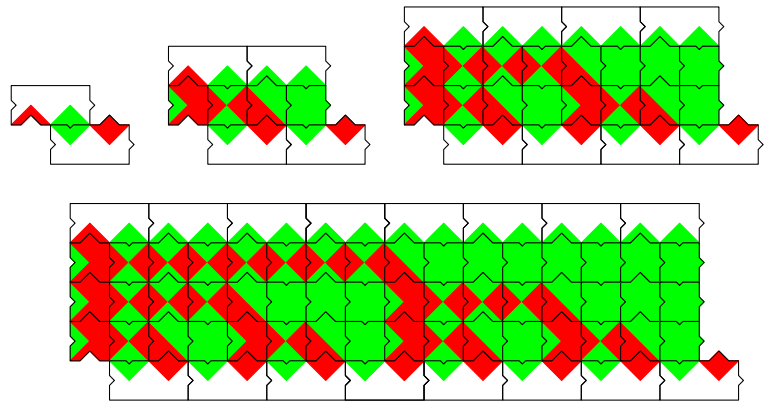
The patterns of usage produced by these rules may be divided into two rough categories—those that reduce their usage of one kind of boundary tile linearly (or slightly faster than linearly) as crystal width grows linearly, and those that reduce their usage of one kind of boundary tile exponentially as crystal width grows linearly. This difference is directly related to cycle size; most of the rules surveyed have a cycle with the same qualitative sort of pattern that grows longer as crystal width increases. The cycles in the rules diagrammed in Figure 7.3 usually use one or a small fixed number of the tile to be restricted in each cycle, so that the per row usage of that tile shrinks as cycle size grows. Examples of the repeat units for two such cycle types are shown in Figure 7.4(a-b).

Provided that the concentration of one type of edge tile is sufficiently small, with each increase in width the tile sets used to make the scatter plots in Figures 7.3b and 7.3c produce a faster-growing crystal type. Despite the fact that our search selected only tile sets for which with every increase in width, a cycle that uses a different percentage of top and bottom tile types appears, a search over the 100 randomly chosen rules turned up 7 rules for which cycles that use fewer of a particular kind of boundary tile than thinner crystals existed for some widths, but not others. Figure 7.3d shows the attractors produced by two such rules and Figure 7.4 shows the cycles produced by these rules. The first rule produces a potentially fitter assembly type with assemblies of 1, 3, 5, 7, and 9 rows of rule tiles, and the second produces such new cycles with assemblies of 1, 2, 4, 5, and 10 tiles. Thus, it would be expected that a more lenient search would turn up other tile sets that could potentially evolve complex crystals when the concentration of some tile types is low. However, evolution of a tile set for which only some widths produce selective advantage would be expected to occur more slowly.

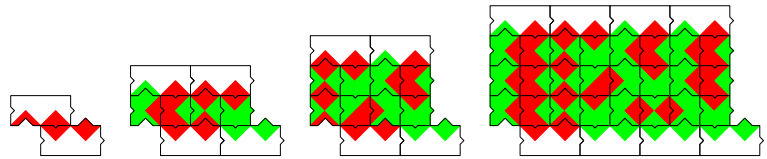
Do such rule sets actually induce evolution toward wider, more complex assemblies in a setting where tile attachment is reversible and stochastic? One roadblock to thinking about this question clearly is that ribbons encoding some attractor cycles can only grow forwards, but others might grow almost reversibly. All things being equal, the latter group would clearly grow more quickly, but the consideration the usage of the different kinds of boundary tiles as growth proceeds in one



a Ribbons from a tile set where ribbons decrease their per-row use of the red bottom tile linearly as width increases.



b Ribbons from a tile set where ribbons decrease their per-row usage of the red top tile exponentially as width increases. Note that the rule tiles in the patterns produced by the tiles in this rule count in binary [CRW04]



c Ribbons from a tile set where ribbons decrease their per-row usage of the red (input) bottom tile linearly with every other width increase

Figure 7.4: **Ribbons from three tile sets where wider ribbons use fewer of a designated tile**

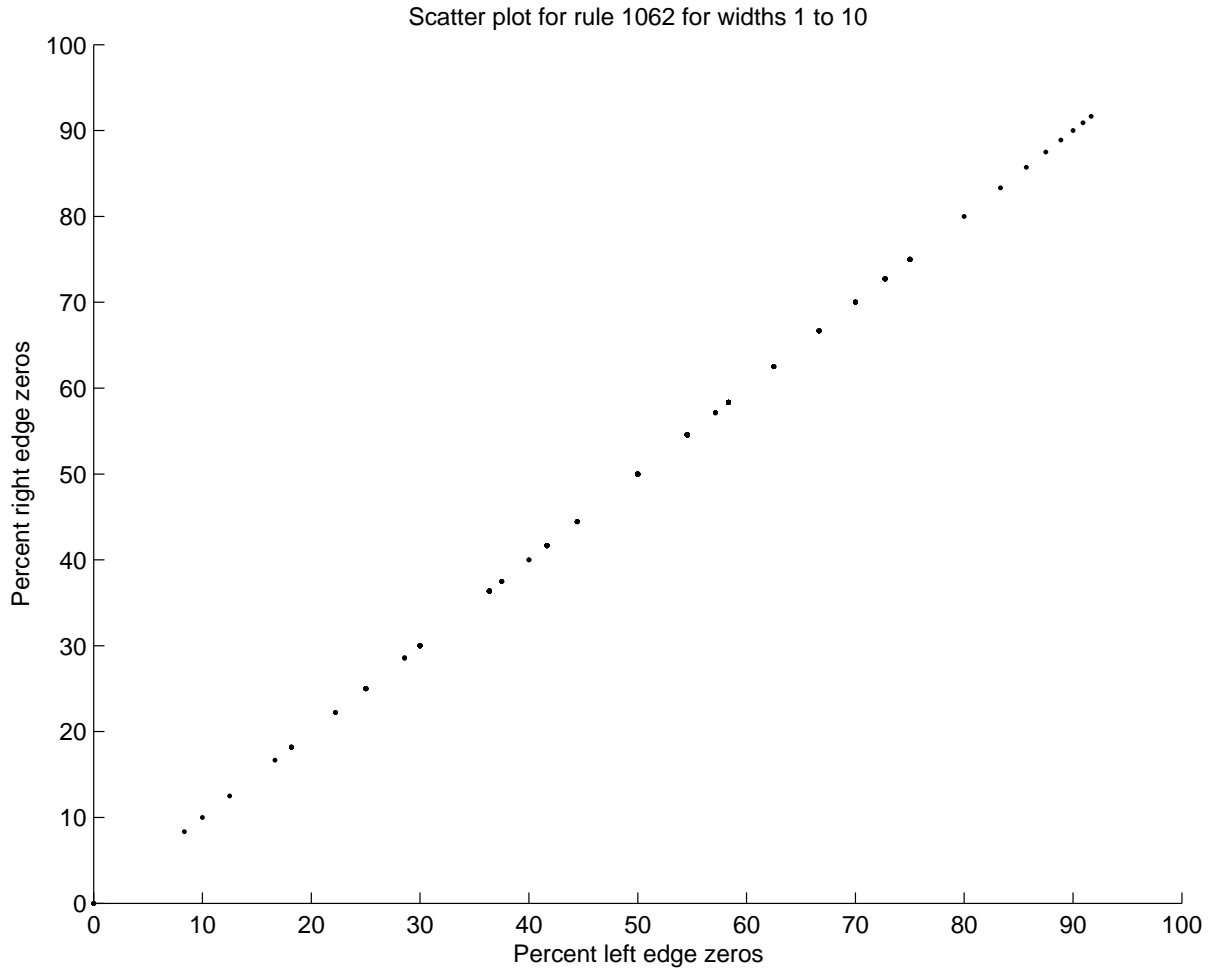


Figure 7.5: **Boundary tile usage by a sample reversible zig-zag tile set**

direction only does not take this factor into account.

In the next section I consider a set of rules for which this is not a concern: all patterns grow equally quickly in both directions. As I will see, there also are rules within this subset of tile sets for which restriction of some kinds of boundary tiles induces evolution of wider crystals.

7.5 Evolution of Zig-Zag Crystals Encoding Reversible Binary Cellular Automata

I better understand the rate of growth of assemblies constructed from reversible zig-zag cellular automata tiles sets. A reversible tile set contains exactly one input for each output so that growth in the reverse direction is deterministic. An example of such a tile set is shown in Figure 7.2a.

There are just 2304 types of reversible binary zig-zag cellular automata, so it is possible to enumerate exhaustively the set of rules which use fewer of some kinds of boundary tiles as they grow wider. A search analogous to the one described in the previous section turned up 16 rules

for which increasing width reduces the rate at which one kind of boundary tiles on each side is used. One example is shown in Figure 7.5. All of the rules that were discovered were similar in important ways. The size of the largest cycle at each width w contains $2(w + 1)$ columns, and with each increasing width, a rule uses 1 of one kind of boundary tile and the rest of the other kind of boundary tile on both the top and the bottom of the ribbon. Since computation is reversible, all possible states are part of a cycle, so that the number of possible cycles for each width increases exponentially. Since the cycle size is small compared to the total number of states, ribbon growth is susceptible to growth errors and most mistakes will change the pattern being copied. (This is in contrast to most of the reversible rules investigated, where there was only one cycle per width. Mistakes in copying these rules would perpetuate the copying of the same pattern.)

7.6 Simulation

To determine whether the proposed selection pressure produces wider ribbons in a more realistic assembly model, I used a stochastic kinetic simulator to track the evolution of crystals with two reversible zig-zag cellular automata tile sets. In both simulations, one of the boundary tiles was available in much lower concentrations than the others. The first tile set was the tile set shown in Figure 7.2a. The tiles from this tile set were such that wider crystals could use progressively fewer of the rare tile to grow. The second tile set was exactly like the first, except that the outputs of the top 2 tiles were swapped. In the latter tile set, wider crystals did not use fewer of the rare tile.

For our simulations, I augmented a previously designed stochastic kinetic simulation of DNA tile assembly [Win98]. The simulation allows tiles to attach to each other or to existing assemblies with a diffusion dependent forward rate ($k_f = 10^6/M/s$ [Wet91]) and a backward rate set by the ΔG of tile attachment, which was assumed to be strictly cooperative: $\Delta G^\circ = 18$ kcal/mol for an attachment by two bonds and $\Delta G^\circ = 9$ for attachment by one bond. The concentration of free tiles was held constant although the concentration of assemblies was allowed to increase. Breakage of assemblies occurred with a small probability per time step, so that a row was sheared on average every 10,000 seconds. Wider assemblies were broken at a slightly smaller rate than thinner ones. Rule tiles were present at $2.25 \mu\text{M}$, three of the four boundary tiles were present at $.45 \mu\text{M}$, and the bottom “1” input tile was present at $.045 \mu\text{M}$.

Figure 7.6 shows the rate at which assemblies of different widths arise in the two simulations. While the total concentration of all assemblies grows at similar rates for both rules (although the time of onset of nucleation is different in the two simulations), the widths of the crystals that grow in the two simulations are completely different. In the first case, where the rare bottom boundary tile is used less often in some kinds of wider assemblies, there is a preference for assemblies three and four tiles wide, the widest kinds I tracked. In the second case, where wider assemblies do not use fewer of the rare bottom tile, there is a preference for the thinnest possible assembly, which is two tiles wide. Thus, stochastic simulations support, to a very limited extent, the rough analysis in the previous sections.

7.7 Conclusions and Open Questions

In this work, I’ve suggested that allowing crystals to grow under conditions in which the concentration of one monomer type is much smaller than the others can induce non-trivial evolution. My investigations were restricted to reducing the concentration of a particular kind of zig-zag crystal

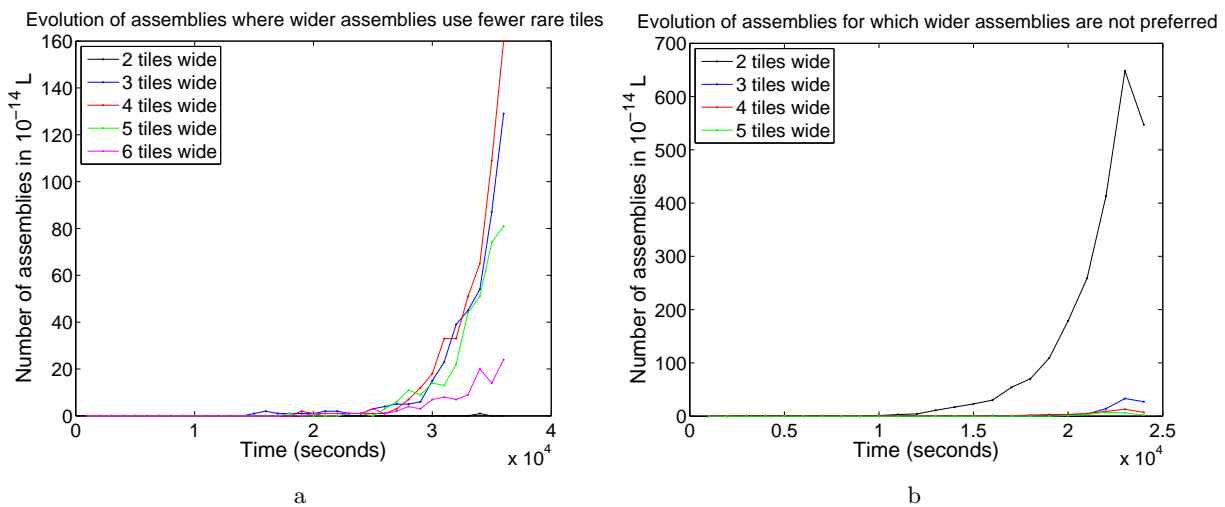


Figure 7.6: Simulations of growth of zig-zag cellular automata tile sets using the kinetic Tile Assembly Model

which copies patterns of zeros and ones. I used analysis and simulation to demonstrate that this evolution occurs in the cases I have discussed. I also provided at least some suggestion that this particular change in concentration is required; at least one other set of concentrations in simulation failed to show evolution of new crystal forms. However, these results are preliminary. Before doing experiments to test whether the phenomena described here can also be seen in the laboratory, we need to further investigate the extent to which these results hold up in a realistic model of DNA tile assembly. In practice we cannot precisely control physical conditions such as temperature or tile concentration. Therefore, we also need to understand how robust these results are with respect to changes in these parameters.

It is particularly important to understand to what extent mistakes in copying patterns and nucleation of new patterns would affect the evolutionary process. Here, I do not consider how the mutation rate would effect the outcome of evolution, but previous work shows that the mutation rate affects any evolutionary process. If mutations occur at more than a certain frequency, evolution becomes impossible [EMS88]. Fortunately, there is reason to believe that some of the tile sets that we study may be robust to many errors and therefore, that they can evolve even under imperfect assembly conditions. For example each width of crystal can copy only one pattern in most of the irreversible cellular automata tile sets that we investigate. A mismatch error in such a tile set would commence growth that assembled a different part of the pattern but would eventually return to the location in the pattern assembly was at before the mismatch. Thus, such errors would not significantly affect the growth rate of the crystal.

Unfortunately, though, the highest acceptable mutation rate for copying and therefore selecting for a pattern decreases with crystal width. Might we as a consequence expect a limit to the sizes of patterns that can be copied under attainable physical conditions? Naively, the answer to this question seems to be yes. But it may be that for some tile sets, as crystals grow wider, their robustness to errors increases. This is true if the number of patterns a tile set can copy grows sub-exponentially with crystal width, except for one small problem: the likelihood of errors that make crystal width smaller increases with width, because at each spot where a rule tile can attach, an edge tile might attach instead. We might hope to design a tile set for which changes in width are particularly penalized—any early attachment of an edge tile would be sure to start copying a particularly unfit pattern—but this seems difficult. We suggest that the concentration of edge tiles in our experiments be smaller than the concentration of rule tile types, so that attachment of a mismatching edge tile instead of a matching edge tile should not occur often.

If we can design tile sets that are robust to mismatch errors and possibly to width change errors, we might also hope to design tile sets that are more robust to other errors. Lattice defect errors, where a column of tiles either “disappears” or “appears” at a wrinkling of the crystal, might be avoided in the same way that we might avoid errors which cause changes in width. Because cellular automaton computation is known to be more robust to noise in 3 dimensions than in 2 [GR88], we might also eventually hope to improve robustness by constructing three dimensional crystals of the kind studied here [CG07]. Particularly tantalizing is the idea that if a crystal suffers an error of any kind that makes it less fit, other layers of the 3-D crystal might simply grow more quickly and even grow around the mistake, preventing it from propagating.

We’ve shown here that crystals can adapt to the rarity of a particular edge tile by growing large and complex in a particular way. By converse, if the concentration of that tile is increased, would crystals become thinner? We’ve studied the effects of reducing the concentration of just one edge tile type, but are there interesting effects if we changed the concentration of multiple edge tile

types or if we changed the concentration of rule tiles in addition to edge tiles? One particularly compelling result might be if we could locate one tile set that could adapt differently to a variety of changes in tile concentration—growing one kind of wide crystals under one set of concentration, thin crystals under another, and a different set of wider crystals under a third set of concentrations. We might hope to watch these different adaptation processes in the lab.

That said, a tile set that adapts to any computable set of environmental conditions has been described previously [SWa], but it was far too complex to synthesize. Thus, we might still be interested in looking for a small set of tiles with the capacity for adaptation. Even the tile set we developed here is somewhat large—we chose it because it was experimentally viable. Are there smaller tile sets that can adapt to tile concentrations by growing more complex crystals? Is there a smallest such tile set? While the question of the smallest tile set with the capacity for evolution seems like a fundamental one, it requires first that we rigorously define what the capacity for evolution is.

And might another kind of simple crystal be capable of evolution? There is no reason to believe that the results described here are particular to the morphology we used. We might imagine tile sets that formed not ribbon crystals but DNA nanotubes [RENP⁺04], 2-D lattices [WLWS98], or 3-D rectangular prisms that copied a 2-D pattern that also have a similar capacity for evolution. We might also imagine it in other kinds of organic crystals such as 3-D protein crystals or prions. If such evolution is possible with a particularly simple tile set, might it also occur in natural crystals such as clays? Many clays use multiple monomer types and have particular constraints on their morphology [Meu05], so such a question is not entirely far-fetched. To consider the possibility that evolution of the kind described here occurs in natural crystals, we might want to understand whether our results still hold in a model of crystal growth that more accurately captures the growth process of these crystals. Such an investigation seems worth pursuing: With reference to the origin of life, clay crystal growth is of particular interest. Also of interest to the origin of life might be whether aggregates of materials that don't have crystalline order might evolve because of a lack of certain materials that could be used during growth. If the probability that materials are added to an aggregate is dependent on the aggregate's composition in whole or part we might imagine that such an effect could be seen.

The idea that even aggregates could evolve because of a need for resources is particularly plausible because the scarcity of resources, whether because of environmental changes or competition, is widely acknowledged as a driving force behind evolution. Evolution because of resource restrictions has been observed in the evolution of artificial computer programs [CWO⁺04], nucleic acid evolution [RJ02, Joy04], and widely in ecology. Thus, it might be that some candidates for the origin of life such as auto-catalytic cycles might also be capable of evolution under the right physical conditions. This, however, remains to be investigated.

The general lesson of this work is in some sense how the power of computational and analytical investigation of a system with the capacity for self-replication and evolution can provide ideas about how non-trivial evolution might occur, and in our case, guide experiments. Thus, it seems possible to attack the mystery of how complex, open-ended evolution might occur with both theory and experiment.

Bibliography

- [ACGH01] Leonard M. Adleman, Qi Cheng, Ashish Goel, and Ming-Deh Huang. *Running time and program size for self-assembled squares*, pages 740–748. ACM, New York, 2001.
- [Ada98] Christoph Adami. *Introduction to Artificial Life*. Springer, 1998.
- [Adl94] Leonard M. Adleman. Molecular computation of solutions to combinatorial problems. *Science*, 266:1021–1024, November 11, 1994.
- [AG04] Gonen Ashkenasy and M. Reza Ghadiri. Boolean logic functions of a synthetic peptide network. *Journal of the American Chemical Society*, 126(36):11140–11141, 2004.
- [AGKS04] Gagan Aggarwal, Michael H. Goldwasser, Ming-Yang Kao, and Robert T. Schweller. *Complexities for generalized models of self-assembly*, pages 880–889. AMS/SIAM, Providence, RI, 2004.
- [AH02] P. Aldrige and K. T. Hughes. Regulation of flagellar assembly. *Current Opinion in Microbiology*, 5(2):160–165, 2002.
- [AOT⁺99] Islamshah Amlani, Alexei O. Orlov, Geza Toth, Gary H. Bernstein, Craig S. Lent, and Gregory L. Snider. Digital logic gate using quantum-dot cellular automata. *Science*, 284(5412):289–291, 1999.
- [Bar07] Robert Barish. Personal communication, 2007.
- [BCT00] Victor A. Bloomfield, Donald M. Crothers, and Ignacio Tinoco, Jr. *Nucleic Acids: Structures, Properties, and Functions*. University Science Books, 2000.
- [Bed07] Mark Bedau. Personal Communication, 2007.
- [Ber66] Robert Berger. The undecidability of the domino problem. *Memoirs of the AMS*, 66:1–72, 1966.
- [BGBD⁺04] Yaakov Benenson, Binyamin Gil, Uri Ben-Dor, Rivka Adar, and Ehud Shapiro. An autonomous molecular computer for logical control of gene expression. *Nature*, 429:423–429, 2004.
- [BH00] J. L. Beuchat and J. O. Haenni. *Von Neumann’s 29-state cellular automaton: a hardware implementation*, volume 43, pages 300–308. IEEE, 2000.

- [BL85] Charles H. Bennett and R. Landauer. Fundamental physical limits of computation. *Scientific American*, 253(1):48–56, 1985.
- [Blu67] M Blum. A machine-independent theory of the complexity of recursive functions. *Journal of the ACM*, 14:322–336, 1967.
- [BMP⁺00] Mark A. Bedau, John S. McCaskill, Norman H. Packard, Steen Rasmussen, Chris Adami, David G. Green, Takashi Ikegami, Kunihiko Kaneko, and Thomas S. Ray. *Open problems in artificial life*, pages 363–376. MIT Press, 2000.
- [Bra95] D. Bray. Protein molecules as computational elements in living cells. *Nature*, 376:307–312, 1995.
- [BRC04] Steven A. Benner, Alonso Ricardo, and Matthew A. Carrigan. Is there a common chemical model for life in the universe? *Current Opinion in Chemical Biology*, 8:672–689, 2004.
- [BRW05] Robert D. Barish, Paul W. K. Rothmund, and Erik Winfree. Two computational primitives for algorithmic self-assembly: Copying and counting. *Nano Letters*, 5:2586–2592, 2005.
- [BS93] D. P. Bartel and J. W. Szostak. Isolation of new ribozymes from a large pool of random sequences. *Science*, 261:1411–1418, 1993.
- [BSRW07] Robert Barish, Rebecca Schulman, Paul W. K. Rothmund, and Erik Winfree. Programming the morphology of DNA-based heterocrystals using a nucleus-encoded bitstring. In preparation, 2007.
- [BTCW97] N. Bowden, A. Terfort, J. Carbeck, and G.M. Whitesides. Self-assembly of mesoscale objects into ordered two-dimensional arrays. *Science*, 276:233–235, 1997.
- [BZTL02] Roy Bar-Ziv, Tsvi Tlusty, and Albert Libchaber. Protein-DNA computation by stochastic assembly cascade. *Proceedings of the National Academy of Sciences USA*, 99(18):11589–11592, 2002.
- [Car06] Sean B. Carroll. *Endless Forms Most Beautiful: The New Science of Evo Devo*. W. W. Norton, 2006.
- [CBG⁺04] Nickolas Chelyapov, Yuriy Brun, Manoj Gopalkrishnan, Dustin Reishus, Bilal Shaw, and Leonard Adleman. DNA triangles and self-assembled hexagonal tilings. *Journal of the American Chemical Society*, 126(43):13924–13925, 2004.
- [CBLS97] Alberto Credi, Vincenzo Balzani, Steven J. Langford, and J. Fraser Stoddart. Logic operations at the molecular level. an XOR gate based on a molecular machine. *Journal of the American Chemical Society*, 119(11):2679–2681, 1997.
- [CDVW04] Sean R. Collins, Adam Douglass, Ronald D. Vale, and Jonathan S. Weissman. Mechanism of prion propagation: Amyloid growth occurs by monomer addition. *PLoS Biology*, 2:e321, 2004.

- [CG05] Ho-Lin Chen and Ashish Goel. Error free self-assembly using error prone tiles. In Ferretti et al. [FMZ05], pages 62–75.
- [CG07] Ho-Lin Chen and Ashish Goel. Submitted, 2007.
- [CJAG95] Gérard Manhès Claude J. Allègre and Christa Göpel. The age of the earth. *Geochimica et Cosmochimica Acta*, 59:1445–1456, 1995.
- [CK06] In-Geol Choi and Sung-Hou Kim. Evolution of protein structural classes and protein sequence families. *Proceedings of the National Academy of Sciences USA*, 103(38):14056–14061, 2006.
- [CODS04] James M. Carothers, Stephane C. Oestreich, Jonathan H. Davis, and Jack W. Szostak. Informational complexity and functional activity of RNA structures. *Journal of the American Chemical Society*, 126:5130–5137, 2004.
- [Coo04] Matthew Cook. Universality in elementary cellular automata. *Complex Systems*, 15(1):1–40, 2004.
- [CR04] Junghuei Chen and John Reif, editors. *DNA Computing 9*, volume LNCS 2943, Berlin Heidelberg, 2004. Springer-Verlag.
- [Cri68] Francis H. C. Crick. The origin of the genetic code. *Journal of Molecular Biology*, 38:367–379, 1968.
- [CRS04] Irene A. Chen, Richard W. Roberts, and Jack W. Szostak. The emergence of competition between model protocells. *Science*, 305:1474–1476, 2004.
- [CRW04] Matthew Cook, Paul W. K. Rothmund, and Erik Winfree. Self-assembled circuit patterns. In Chen and Reif [CR04], pages 91–107.
- [CS66] A. Graham Cairns-Smith. The origin of life and the nature of the primitive gene. *Journal of Theoretical Biology*, 10:53–88, 1966.
- [CS82] A. Graham Cairns-Smith. *Genetic Takeover and the Mineral Origins of Life*. Cambridge University Press, 1982.
- [CS88] A. Graham Cairns-Smith. The chemistry of materials for artificial Darwinian systems. *International Reviews in Physical Chemistry*, 7:209–250, 1988.
- [CS06] A. Graham Cairns-Smith. Personal communication, 2006.
- [CSGW07] Ho-Lin Chen, Rebecca Schulman, Ashish Goel, and Erik Winfree. Preventing facet nucleation during algorithmic self-assembly. Submitted, 2007.
- [CSH86] A. Graham Cairns-Smith and Hyman Hartman. *Clay Minerals and the Origin of Life*. Cambridge University Press, 1986.
- [CSK⁺04] Arkadiusz Chworos, Isil Severcan, Alexey Y. Koyfman, Patrick Weinkam, Emin Oroudjev, Helen G. Hansma, and Luc Jaeger. Building programmable jigsaw puzzles with RNA. *Science*, 306:2068–2072, 2004.

- [CSWB07] Matthew Cook, David Soloveichik, Erik Winfree, and Jehoshua Bruck. In preparation, 2007.
- [CWO⁺04] Stephanie S. Chow, Claus O. Wilke, Charles Ofria, Richard Lenski, and Christoph Adami. Adaptive radiation from resource competition in digital organisms. *Science*, 305:84–86, 2004.
- [Dar88] Francis Darwin, editor. *The Life and Letters of Charles Darwin*, volume III. John Murray, London, 1888.
- [Dav01] Eric H. Davidson. *Genomic Regulatory Systems: Development and Evolution*. Academic Press, 2001.
- [Dav06] Eric H. Davidson. *The Regulatory Genome: Gene Regulatory Networks In Development And Evolution*. Academic Press, 2006.
- [DB02] Ken A. Dill and Sarina Bromberg. *Molecular Driving Forces: Statistical Thermodynamics in Chemistry and Biology*. Garland Science, New York, 2002.
- [DG00] Roger Davey and John Garside. *From Molecules to Crystallizers*. Oxford University Press, Oxford, UK, 2000.
- [DLWP04] Robert M. Dirks, Milo Lin, Erik Winfree, and Niles A. Pierce. Paradigms for computational nucleic acid design. *Nucleic Acids Research*, 32:1392, 2004.
- [DM97] Arshad Desai and Timothy J. Mitchison. Microtubule polymerization dynamics. *Annual Review of Cell Developmental Biology*, 13:83–117, 1997.
- [Dob64] Theodosius Dobzhansky. Biology, molecular and organismic. *American Zoologist*, 4:443–452, 1964.
- [DP04] Robert M. Dirks and Niles A. Pierce. Triggered amplification by hybridization chain reaction. *Proceedings of the National Academy of Sciences USA*, 101(43):15275–15278, 2004.
- [DS86] Marian C. Diamond and Arnold B. Scheibel. *The Human Brain Coloring Book*. HarperCollins, New York, 1986.
- [EG00] Brent Ewing and Phil Green. Analysis of expressed sequence tags indicates 35,000 human genes. *Nature Genetics*, 25:232–234, 2000.
- [Eig71] Manfred Eigen. Self-organization of matter and evolution of biological macromolecules. *Naturwissenschaften*, 58(10):465–523, 1971.
- [EMS88] Manfred Eigen, John McCaskill, and Peter Schuster. Molecular quasi-species. *Journal of Physical Chemistry*, 92:6881–6891, 1988.
- [ENKF04] Axel Ekani-Nkodo, Ashish Kumar, and D. Kuchnir Fygenon. Joining and scission in the self-assembly of nanotubes from DNA tiles. *Physical Review Letters*, 93:268301, 2004.

- [FC06] Anthony C. Forster and George M. Church. Towards synthesis of a minimal cell. *Molecular Systems Biology*, 2, 2006.
- [Fel92] David A. Fell. Metabolic control analysis: a survey of its theoretical and experimental development. *Biochemical Journal*, 286:313–330, 1992.
- [FFLN77] Arlette Fellous, Jacques Francon, Ana-Maria Lennon, and Jacques Nunez. Microtubule assembly *in vitro*. *European Journal of Biochemistry*, 78:167–174, 1977.
- [FFS⁺95] D. Kuchnir Fygenson, H. Flyvbjerg, K. Sneppen, A. Libchaber, and S. Leibler. Spontaneous nucleation of microtubules. *Physical Review E*, 51:5058–5063, 1995.
- [FHP⁺07] Kenichi Fujubayashi, Rizal Hariadi, Sung-Ha Park, Erik Winfree, and Satoshi Murata. Toward reliable algorithmic self-assembly of DNA tiles: a fixed-width cellular automaton pattern. In preparation, 2007.
- [FKP86] J. Doyne Farmer, Stuart A. Kauffman, and Norman H. Packard. Autocatalytic replication of polymers. *Physica D*, 22:50–67, 1986.
- [FM04] Kenichi Fujibayashi and Satoshi Murata. *A Method of Error Suppression for Self-Assembling DNA Tiles*. Springer-Verlag, Berlin Heidelberg, 2004.
- [FMZ05] Claudio Ferretti, Giancarlo Mauri, and Claudio Zandron, editors. *DNA Computing 10*, volume LNCS 3384, Berlin Heidelberg, 2005. Springer-Verlag.
- [Fon92] Walter Fontana. Algorithmic chemistry. In Christopher G. Langton, Charles Taylor, J. Doyne Farmer, and Steen Rasmussen, editors, *Artificial Life II*, pages 159–211. Addison-Wesley, 1992.
- [FPU55] Enrico Fermi, John Pasta, and Stanslav Ulam. Studies of nonlinear problems I. Technical Report LA-1940, Los Alamos Scientific Library, 1955.
- [FS93] Tsu-Ju Fu and Nadrian C. Seeman. DNA double-crossover molecules. *Biochemistry*, 32:3211–3220, 1993.
- [FS98] Walter Fontana and Peter Schuster. Shaping space: the possible and the attainable in RNA genotype. *Journal of Theoretical Biology*, 194:491–515, 1998.
- [GC04] Indraneel Ghosh and Jean Chmielewski. Peptide self-assembly as a model of proteins in the pre-genomic world. *Current Opinion in Chemical Biology*, 8:640–644, 2004.
- [GGS05] Jana Gavertz, Hin Hark Gan, and Tamar Schlick. In vitro RNA random pools are not structurally diverse: A computational analysis. *RNA*, 11:853–863, 2005.
- [Gil76] D. T. Gillespie. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *J. Comput. Phys.*, 22:403–434, 1976.
- [Gol94] Solomon W. Golomb. *Polyominoes*. Princeton University Press, Princeton, N.J., 2nd edition, 1994.
- [GR88] Peter Gács and John Reif. A simple three-dimensional real-time reliable cellular array. *Journal of Computer and System Sciences*, 36(2):125–147, 1988.

- [Har04] Rizal Hariadi. Personal communication, 2004.
- [HCL⁺05] Yu He, Yi Chen, Haipeng Liu, Alexander E. Ribbe, and Chengde Mao. Self-assembly of hexagonal DNA two-dimensional (2D) arrays. *Journal of the American Chemical Society*, 127:12202–12203, 2005.
- [HFS03] M. M. Hanczyc, S. M. Fujikawa, and J. W. Szostak. Experimental models of primitive cellular compartments: Encapsulation, growth, and division. *Science*, 302:618–622, 2003.
- [HS96] Liisa Holm and Chris Sander. Mapping the protein universe. *Nature*, 273(5275):595–602, 1996.
- [HW] Claudia Huber and Günter Wächtershäuser. Activated acetic acid by carbon fixation on (Fe,Ni)S under primordial conditions. *Science*, 276(5210):245–247.
- [Int07] Intel microprocessor quick reference guide. Technical report, 2007. <http://www.intel.com/pressroom/kits/quickreffam.htm>.
- [Joy87] Gerald F. Joyce. Nonenzymatic template-directed synthesis of informational macromolecules. In *Cold Spring Harbor Symposia on Quantitative Biology*, volume 52, pages 41–51, 1987.
- [Joy89] Gerald F. Joyce. Amplification, mutation and selection of catalytic RNA. *Gene*, 82:83–87, 1989.
- [Joy96] Gerald F. Joyce. Ribozymes: Building the RNA world. *Current Biology*, 6:965–967, 1996.
- [Joy04] Gerald F. Joyce. Directed evolution of nucleic acid enzymes. *Annual Review of Biochemistry*, 73:791–836, 2004.
- [JUL⁺01] Wendy K. Johnston, Peter J. Unrau, Michael S. Lawrence, Margaret E. Glasner, and David P. Bartel. RNA-catalyzed RNA polymerization: Accurate and generate RNA-templated primer extension. *Science*, 292:1319–1325, 2001.
- [KAP95] J. F. Kelleher, S. J. Atkinson, and T. D. Pollard. Sequences, structural models and cellular localization of the actin-related proteins. *Journal of Cell Biology*, 10:197–210, 1995.
- [Kau93] Stuart A. Kauffman. *The Origins of Order*. Oxford University Press, 1993.
- [LB96] Richard J. Lipton and Eric B. Baum, editors. *DNA Based Computers*, volume 27 of *DIMACS*, Providence, RI, 1996. AMS.
- [LCMY03] Catherine Lozupone, Shankar Changayil, Irene Majerfeld, and Michael Yarus. Selection of the simplest RNA that binds isoleucine. *RNA*, 9:1315–1322, 2003.
- [LE03] M. Levy and Andrew D. Ellington. Exponential growth by cross-catalytic cleavage of deoxyribozymogens. *Proceedings of the National Academy of Sciences USA*, 100:6416–6421, 2003.

- [Leh93] Jean-Marie Lehn. Supramolecular chemistry. *Science*, 260:1762–1764, 1993.
- [Lev73] L. A. Levin. Universal sequential search problems. *Problems of Information Transmission*, 9:265–266, 1973.
- [LGMKs96] David H. Lee, Juan R. Granja, Jose A. Martinez, and M. Reza Ghadiri Kay Severi and. A self-replicating peptide. *Nature*, 382(6591):525–528, 1996.
- [LK97] Andrea C. Levi and Miroslav Kotrla. Theory and simulation of crystal growth. *Journal of Physics: Condensed Matter*, 9:299–344, 1997.
- [LL00] Michail G. Lagoudakis and Thomas H. LaBean. 2-D DNA self-assembly for satisfiability. In Erik Winfree and David K. Gifford, editors, *DNA Based Computers V*, volume 54 of *DIMACS*, pages 141–154, Providence, RI, 2000. AMS.
- [LPRY04] Dage Liu, Sung Ha Park, John H. Reif, and Hao Yan. DNA nanotubes self-assembled from triple-crossover tiles as template for conductive nanowires. *Proceedings of the National Academy of Sciences USA*, 101:717–722, 2004.
- [LV97] Ming Li and Paul Vitanyi. *An Introduction to Kolmogorov Complexity and Its Applications (Second Edition)*. Springer Verlag, New York, 1997.
- [LYK⁺00] Thomas H. LaBean, Hao Yan, Jens Kopatsch, Furong Liu, Erik Winfree, John H. Reif, and Nadrian C. Seeman. Construction, analysis, ligation and self-assembly of DNA triple crossover complexes. *Journal of the American Chemical Society*, 122(9):1848–1860, 2000.
- [LYQS96] Xiaojun Li, Xiaoping Yang, Jing Qi, and Nadrian C. Seeman. Antiparallel DNA double crossover molecules as components for nanoconstruction. *Journal of the American Chemical Society*, 118(26):6131–6140, 1996.
- [Mag97] Marcelo O. Magnasco. Chemical kinetics is Turing universal. *Physical Review Letters*, 78:1190–1193, 1997.
- [MAM⁺96] S. J. Mojzsis, G. Arrhenius, K. D. McKeegan, T. M. Harrison, A. P. Nutman, and C. R. L. Friend. Evidence for life on earth before 3,800 million years ago. *Nature*, 384:55–59, 1996.
- [Man88] Stephen Mann. Molecular recognition in biomineralization. *Nature*, 332:119–124, 1988.
- [Mar03] Ivan V. Markov. *Crystal Growth for Beginners*. World Scientific, Singapore, 2003.
- [MB87] L. A. Marky and K. J. Breslauer. Calculating thermodynamic data for transitions of any molecularity from equilibrium melting curves. *Biopolymers*, 26(9):1601–1620, 1987.
- [MBS⁺95] Michelle Moritz, Michael B. Braunfeld, John W. Sedat, Bruce Alberts, and David A. Agard. Microtubule nucleation by big gamma-tubulin-containing rings in the centrosome. *Nature*, 378:638–640, 1995.

- [McD62] J.E. McDonald. Homogenous nucleation of vapor condensation. *American Journal of Physics*, 30:870, 1962.
- [Meu05] Alain Meunier. *Clays*. Springer, 2005.
- [MFH92] Melanie Mitchell, Stephanie Forrest, and John H. Holland. The royal road for genetic algorithms: Fitness landscapes and GA performance. In *Proceedings of the First European Conference on Artificial Life*, 1992.
- [MHM⁺04] James C. Mitchell, J. Robin Harris, Jonathan Malo, Jonathan Bath, and Andrew J. Turberfield. Self-assembly of chiral DNA nanotubes. *Journal of the American Chemical Society*, 126(50):15342–16343, 2004.
- [Mil53] Stanley L. Miller. A production of amino acids under possible primitive earth conditions. *Science*, 117:528–529, 1953.
- [MLK⁺05] F. Mathieu, S. P. Liao, J. Kopatscht, T. Wang, C. D. Mao, and N. C. Seeman. Six-helix bundles designed from DNA. *Nano Letters*, 5:661–665, 2005.
- [MLRS00a] Chengde Mao, Thomas H. LaBean, John H. Reif, and Nadrian C. Seeman. Logical computation using algorithmic self-assembly of DNA triple-crossover molecules. *Nature*, 407(6803):493–496, 2000.
- [MLRS00b] Chengde Mao, Thomas H. LaBean, John H. Reif, and Nadrian C. Seeman. Logical computation using algorithmic self-assembly of DNA triple-crossover molecules. *Nature*, 407(6803):493–496, 2000.
- [MSS99] Chengde Mao, Weiqiong Sun, and Nadrian C. Seeman. Designed two-dimensional DNA Holliday junction arrays visualized by atomic force microscopy. *Journal of the American Chemical Society*, 121(23):5437–5443, 1999.
- [Opa03] Aleksandr Oparin. *Origin of Life*. Dover Phoenix Books, Mineola, NY, 2003.
- [Org68] Leslie E. Orgel. Evolution of the genetic apparatus. *Journal of Molecular Biology*, 38:381–393, 1968.
- [Org98a] Leslie E. Orgel. The origin of life—how long did it take? *Origins of Life and Evolution of the Biosphere*, 28:91–96, 1998.
- [Org98b] Leslie E. Orgel. The origin of life: A review of facts and speculations. *Trends in Biochemical Sciences*, 23:491–495, 1998.
- [PC86] Thomas D. Pollard and John A. Cooper. Actin and actin-binding proteins: a critical evaluation of mechanisms and functions. *Annual Review of Biochemistry*, 55:987–1035, 1986.
- [PEG⁺95] S. Pitsch, A. Eschenmoser, B. Gedulin, S. Hui, and G. Arrhenius. Mineral induced formation of sugar phosphates. *Origins of Life and Evolution of Biospheres*, 25(4):297–334, 1995.

- [Pen58] Lionel S. Penrose. Mechanics of self-reproduction. *Ann. of Human Genetics*, 23:59–72, 1958.
- [Pes95] U. Pesavento. An implementation of von Neumann’s self-reproducing machine. *Artif. Life*, 2:337–354, 1995.
- [PJ02] Natasha Paul and Gerald F. Joyce. A self-replicating ligase ribozyme. *Proceedings of the National Academy of Sciences USA*, 99:12733–12740, 2002.
- [PP57] L.S. Penrose and Roger Penrose. A self-reproducing analogue. *Nature*, 179(4571):1183, 1957.
- [QW89] Robin S. Quartin and James G. Wetmur. Effect of ionic strength on the hybridization of oliodeoxynucleotides with reduced charge due to methylphosphonate linkages to unmodified oligodeoxynucleotides containing the complementary sequence. *Biochemistry*, 28:1040–1047, 1989.
- [RACP93] James A. Reggia, Steven L. Armentrout, Hui-Hsien Chou, and Yun Peng. Simple systems that exhibit self-directed replication. *Science*, 26(5099):1282–1287, 1993.
- [Rei97] John H. Reif. Local parallel biomolecular computation. In Harvey Rubin and David Harlan Wood, editors, *DNA Based Computers III*, volume 48 of *DIMACS*, pages 217–254, Providence, RI, 1997. American Mathematical Society.
- [RENP+04] Paul W. K. Rothmund, Axel Ekani-Nkodo, Nick Papadakis, Ashish Kumar, Deborah Kuchnir Fygeneseon, and Erik Winfree. Design and characterization of programmable DNA nanotubes. *Journal of the American Chemical Society*, 126(50):16344–16352, 2004.
- [RJ90] Debra L. Robertson and Gerald F. Joyce. Selection in vitro of an RNA enzyme that specifically cleaves single-stranded DNA. *Nature*, 344:467–468, 1990.
- [RJ02] John S. Reader and Gerald F. Joyce. A ribozyme composed of only two different nucleotides. *Nature*, 420:841–844, 2002.
- [Rot96] Paul W. K. Rothmund. A DNA and restriction enzyme implementation of Turing Machines. In Lipton and Baum [LB96], pages 75–119.
- [Rot00] Paul W. K. Rothmund. Using lateral capillary forces to compute by self-assembly. *Proceedings of the National Academy of Sciences USA*, 97:984–989, 2000.
- [Rot01] Paul W. K. Rothmund. *Theory and Experiments in Algorithmic Self-Assembly*. PhD thesis, University of Southern California, Department of Computer Science, 2001.
- [Rot06] Paul W. K. Rothmund. Folding DNA to create nanoscale shapes and patterns. *Nature*, 440:297–302, 2006.
- [RPW04] Paul W. K. Rothmund, Nick Papadakis, and Erik Winfree. Algorithmic self-assembly of DNA Sierpinski triangles. *PLOS Biology*, 2:424–436, 2004.

- [RSSF03] Micheline Fromont Racine, Bruno Sengerb, Cosmin Saveanua, and Franco Fasiolob. Ribosome assembly in eukaryotes. *Gene*, 313:17–42, 2003.
- [RSY05] John H. Reif, Sudheer Sahu, and Peng Yin. Compact error-resilient computational DNA tiling assemblies. In Ferretti et al. [FMZ05], pages 293–307.
- [RW00] Paul W. K. Rothmund and Erik Winfree. The program-size complexity of self-assembled squares. In *Symposium on Theory of Computing (STOC)*, pages 459–468, New York, 2000. ACM.
- [San98] John SantaLucia, Jr. A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proceedings of the National Academy of Sciences USA*, 95:1460–1465, 1998.
- [SBEDL01] Daniel Segré, Dafna Ben-Eli, David W. Deamer, and Doron Lancet. The lipid world. *Origins of Life and Evolution of Biospheres*, 31(1–2):119–145, 2001.
- [SBEL00] Daniel Segré, Dafna Ben-Eli, and Doron Lancet. Compositional genomes: prebiotic information transfer in mutually catalytic non-covalent assemblies. *Proceedings of the National Academy of Sciences USA*, 97(8):4112–4117, 2000.
- [SBL01] Jack W. Szostak, David P. Bartel, and P. Luigi Luisi. Synthesizing life. *Nature*, 409:387–390, 2001.
- [Sch44] Erwin Schrödinger. *What is Life?* Cambridge University Press, Cambridge, 1944.
- [See90] Nadrian C. Seeman. *De novo* design of sequences for nucleic acid structural engineering. *Journal of Biomolecular Structure & Dynamics*, 8(3):573–581, 1990.
- [Sip97] Michael Sipser. *Introduction to the Theory of Computation*. PWS Publishing Company, 1997.
- [Sip98] Moshe Sipper. Fifty years of research on self-replication: An overview. *Artificial Life*, 4:237–257, 1998.
- [SJB99] V. A. Shneidman, K. A. Jackson, and K. M. Beatty. On the applicability of the classical nucleation theory in an Ising system. *Journal of Chemical Physics*, 111:6932–6941, 1999.
- [SKS96] Oliver Steinbock, Petteri Kettunen, and Kenneth Schowalter. Chemical wave logic gates. *Journal of Physical Chemistry*, 100:18970–18975, 1996.
- [SLPW04] Rebecca Schulman, Shaun Lee, Nick Papadakis, and Erik Winfree. One dimensional boundaries for DNA tile self-assembly. In Chen and Reif [CR04], pages 108–125.
- [SM01] David Sept and J. Andrew McCammon. Thermodynamics and kinetics of actin filament nucleation. *Biophysical Journal*, 81:867–874, 2001.
- [Smi70] John Maynard Smith. Natural selection and the concept of a protein space. *Nature*, 225:563–564, 1970.

- [Sol64] R. J. Solomonoff. A formal theory of inductive inference: Part 1 and 2. *Information Control*, 7:1–22, 224–254, 1964.
- [SS95] John Maynard Smith and Eors Szathmary. *The Major Transitions in Evolution*. W.H. Freeman and Co., 1995.
- [SvK98] Dirk Sievers and Günter von Kiedrowski. Self-replication of hexadeoxynucleotide analogues Autocatalysis versus cross-catalysis. *Chemistry—A European Journal*, 4:629–641, 1998.
- [SWa] Rebecca Schulman and Erik Winfree. How crystals that sense and respond to their environment could evolve. *Natural Computing*. To appear.
- [SWb] Rebecca Schulman and Erik Winfree. Self-replication of DNA crystals. Preliminary data.
- [SW04] David Soloveichik and Erik Winfree. Complexity of self-assembled shapes. In *DNA Computing 10*, Berlin Heidelberg, 2004. Springer-Verlag.
- [SW05a] Rebecca Schulman and Erik Winfree. Controlling nucleation rates in algorithmic self-assembly. 2005.
- [SW05b] Rebecca Schulman and Erik Winfree. Self-replication and evolution of DNA crystals. In *Advances in Artificial Life, 8th European Conference*, volume 3630, Berlin Heidelberg, 2005. Springer-Verlag.
- [SW05c] David Soloveichik and Erik Winfree. Complexity of compact proofreading for self-assembled patterns. In *DNA Computing 11*, Berlin Heidelberg, 2005. Springer-Verlag.
- [SW07] Rebecca Schulman and Erik Winfree. Synthesis of crystals with a programmable kinetic barrier to nucleation. Submitted, 2007.
- [TKM02] Kohji Tomit, Haruhisa Kurokawa, and Satoshi Murata. Graph automata: natural expression of self-reproduction. *Physica D*, 171:197–210, 2002.
- [VAM⁺01] Craig Venter, Mark D. Adams, Eugene W. Myers, Peter W. Li, Richard J. Mural, Granger G. Sutton, Hamilton O. Smith, and Mark Yandell et al. The sequence of the human genome. *Science*, 291(5507):1304–1351, 2001.
- [vK86] Gunter von Kiedrowski. A self-replicating hexadeoxynucleotide. *Angewandte Chemie International Edition*, 25:932–935, 1986.
- [vNAWB66] John von Neumann and ed. A. W. Burks. *The Theory of Self-Reproducing Automata*. University of Illinois Press, 1966.
- [Wäc88] Günter Wächtersäuser. Before enzymes and templates: theory of surface metabolism. *Microbiology and Molecular Biology Reviews*, 52(4):452–484, 1988.
- [Wäc90] Günter Wächtersäuser. Evolution of the first metabolic cycles. *Proceedings of the National Academy of Sciences USA*, 87:200–204, 1990.

- [Wan61] Hao Wang. Proving theorems by pattern recognition. II. *Bell System Technical Journal*, 40:1–42, 1961.
- [Wan62] Hao Wang. An unsolvable problem on dominoes. Technical Report BL-30 (II-15), Harvard Computation Laboratory, 1962.
- [Wat61] Shigeru Watanabe. 5-symbol 8-state and 5-symbol 6-state universal turing machines. *Journal of the Association for Computing Machinery*, 8:476–483, 1961.
- [WB04] Erik Winfree and Renat Bekbolatov. Proofreading tile sets: Error-correction for algorithmic self-assembly. In Chen and Reif [CR04], pages 126–144.
- [WBT68] Richard C. Weisenberg, Gary C. Borisy, and Edwin W. Taylor. The colchicine-binding protein of mammalian brain and its relation to microtubules. *Biochemistry*, 7(12):4466–4479, 1968.
- [WC53] James D. Watson and Francis H. Crick. Molecular structure of nucleic acids. *Nature*, 171:737, 1953.
- [Wet91] James G. Wetmur. DNA probes: Applications of the principles of nucleic acid hybridization. *Critical Reviews in Biochemistry and Molecular Biology*, 36:227–259, 1991.
- [Win96] Erik Winfree. On the computational power of DNA annealing and ligation. In Lipton and Baum [LB96], pages 199–221.
- [Win98] Erik Winfree. Simulations of computing by self-assembly. Technical Report CS-TR:1998.22, Caltech, 1998.
- [Win06] Erik Winfree. Self healing tile sets. In Junghuei Chen, Natasha Jonoska, and Grzegorz Rozenberg, editors, *Nanotechnology: Science and Computation*, pages 55–78. Springer, 2006.
- [WJ97] Martin C. Wright and Gerald F. Joyce. Continuous in vitro evolution of catalytic function. *Science*, 276:614–617, 1997.
- [WL92] Andrew Wuensche and Mike Lesser. *The Global Dynamics of Cellular Automata: An Atlas of Basin of Attraction Fields of One-Dimensional Cellular Automata*. Perseus Books, 1992.
- [WLWS98] Erik Winfree, Furong Liu, Lisa A. Wenzler, and Nadrian C. Seeman. Design and self-assembly of two-dimensional DNA crystals. *Nature*, 394:539–544, 1998.
- [WM02] Matthew D. Welch and R. Dyche Mullins. Cellular control of actin nucleation. *Annual Reviews of Cellular and Developmental Biology*, 18:247–288, 2002.
- [WMS91] George M. Whitesides, John P. Mathias, and Christopher T. Seto. Molecular self-assembly and nanochemistry: a chemical strategy for the synthesis of nanostructures. *Science*, 254:1312–1319, 1991.
- [Wol02] Steven Wolfram. *A New Kind of Science*. Wolfram Media, 2002.

- [WWF⁺94] Peter Walde, Roger Wick, Massimo Fresta, Annarosa Mangone, and Pier Luigi Luisi. Autopoetic self-reproduction of fatty acid vesicles. *Journal of the American Chemical Society*, 116:11649–11654, 1994.
- [YHS⁺07] Peng Yin, Rizal F. Hariadi, Sadheer Sahu, H. M. T. Choi, Sung-Ha Park, B. Walters, Thom H. LaBean, and John H. Reif. Crystals assembled from flexible single-strand DNA tiles. In preparation, 2007.
- [YLFR03] Hao Yan, Thomas H. LaBean, Liping Feng, and John H. Reif. Directed nucleation assembly of DNA tile complexes for barcode-patterned lattices. *Proceedings of the National Academy of Sciences USA*, 100(14):8103–8108, 2003.
- [YPF⁺03] Hao Yan, Sung Ha Park, Gleb Finkelstein, John H. Reif, and Thomas H. LaBean. DNA-templated self-assembly of protein arrays and highly conductive nanowires. *Science*, 301:1882–1884, 2003.
- [YWA02] Yohei Yokobayashi, Ron Weiss, and Frances H. Arnold. Directed evolution of a genetic circuit. *Proceedings of the National Academy of Sciences USA*, 99:16587–16591, 2002.
- [Zet69] A. C. Zettlemoyer, editor. *Nucleation*. Marcel Dekker, New York, 1969.