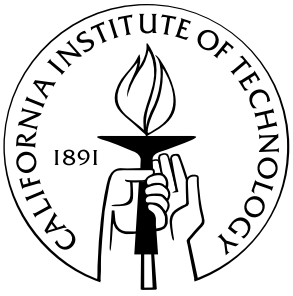
Dynamic DNA Strand Displacement Circuits

Thesis by
David Yu Zhang

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



California Institute of Technology
Pasadena, California
2010
(Defended May 14, 2010)

Acknowledgements

When I started my graduate career–actually, even about one year before I finished my graduate career–I had no idea how anything I did would fit into a grander vision of science and technology. You see, a graduate student's mark of success is somewhat different from a professor's: A grad student's good day is to get his data to agree with his data from yesterday, and to get both to agree with his model (after some model tinkering). A professor's good day is to get his funding agency to agree with his research, and to get the world to agree with his genius. For me, after a somewhat lengthy undergrad detour into transcriptional circuitry work, I had figured out the trick to being a successful graduate student: work with the simplest, most well-understood system possible, and work around anything you don't understand. I think my adviser, Erik Winfree, partially realized what I was doing, and started nagging me from fairly early on to develop a "deeper understanding" of the science that I did, and also to confront rather than avoid some of the harder problems. It wasn't easy (as testified by the 2 and a half years it took to write the manuscript in chapter 2), but in the end I feel more like a scientist than a lab technician because of the experience. So, it is with great humbleness and appreciation that I first give thanks to my Ph.D. adviser, Erik Winfree.

Paul Rothemund, Rebecca Schulman, and Jongmin Kim have provided me with invaluable advice, both scientific and personal, and I feel that I would not be where I am today without their assistance. Paul inspired me to deal with my own adversities with poise by relating to me his past troubles, as well as the ways he dealt with them. Rebecca was and is an endless source of useful information, advice, and encouragement–she was the one who encouraged me to apply for the Hertz despite my undergrad GPA, and she is the one who helped me regain my self confidence in the last few turbulent years of my undergraduate education. Jongmin acted as my SURF co-mentor when I was an undergrad, and taught

me a number of techniques, both in science and in Starcraft. All three of you were great advisers to me, and I am sure all of your future students will feel the same!

I thank my research collaborators, Bernard Yurke, Andrew Turberfield, Georg Seelig, David Soloveichik, Rizal Hariadi, and Lulu Qian. Working with you all has been a wonderful and enlightening experience, and I look forward to future work together.

I thank my other labmates, Nadine Dabby, Joseph Schaeffer, Peng Yin, and Sungwook Woo. Lab would be a much more boring place without your presence.

I give special thanks to Karolyn Knoll, our administrative assistant, for making day-to-day life about a hundred times easier. Thank you Karolyn, for your help ordering reagents and fixing machines and scheduling travel and so on, and thank you for the cookies.

Siping Han, Robert Barish, and Xiaoyan Robert Bao have been three of my best friends during my time here at Caltech, and I thank them for providing not only amusing personal stories, but also intelligent minds that I could bounce ideas off. I am confident that they will become great scientists, and look forward to talking to and working with them in the future.

I thank Niles Pierce, Richard Murray, Zhen-Gang Wang, Scott Fraser, Joel Burdick, and Shuki Bruck, the members of my thesis committee, for their guidance. At the time, I wasn't so happy that they assigned me three whole textbooks to read (and to be honest, I only finished one of them), but in retrospect, I agree that this was important for my development as a scientist.

I am very grateful to the Fannie and John Hertz Foundation, not only for providing me with a nice stipend, but more importantly for providing me with a community of peers who are as excited by science and technology as I am. No other organization I belong to or know of has a stronger sense of unity, and I am proud to be a Hertz Fellow.

My father, mother, and little sister Wendy deserve special thanks for putting up with me as long as they have. I am terrible at keeping in touch at a distance, and it is only through their tireless effort that I still feel that I have a family.

Finally, I would like to thank my fiancee Sherry Xi Chen, who has made me a much happier man in the three years since we met.

Abstract

Nucleic acids, the "NA" in DNA and RNA, have long been known to be vitally important molecules within biological cells and organisms. However, they are interesting for more than just their known roles in biology: their predictable Watson-Crick base pairing properties allow nucleic acids to be powerful nanoscale engineering tools. Additionally, nucleic acid-based devices are particularly attractive as biotechnological tools, because nucleic acids naturally exist within all life, and thus nucleic acid devices more easily function in cellular environments. It is for these reasons that nucleic acids have emerged as a frequent star in recent synthetic biology, biotechnology, and nanotechnology research papers.

This thesis is a collection of 6 experimental papers, 3 theoretical papers, and 1 review paper that demonstrate and characterize novel nucleic acid-based devices such as catalysts, logic gates, and allosteric switches. Particular effort was placed in ensuring that all the designs are generalizable in sequence and that all the devices are modular in nature; this allows many different components to be integrated into higher-complexity devices.

The works presented in this thesis were designed using only non-covalent changes to nucleic acid complexes and structures via Watson-Crick base pairing–i.e. hybridization, branch migration, and dissociation. These three primitives are sufficient to construct an endless variety of circuits and devices, much like how resistors, capacitors, and inductors allow complex electrical circuits. One advantage of devices, reactions, and circuits engineered using only Watson-Crick interactions is their robustness to their environmental conditions. While enzymatic reactions require specific temperatures, salt conditions, and co-factors, nucleic acid hybridization works reliably in a variety of different solutions.

These works are not meant to be final, optimized designs for devices, but rather demonstrations of the wide range of possibilities afforded by nucleic acid engineering and of problems that can be practically solved with dynamic nucleic acid devices in the near future.

Contents

Chapter 1: DNA as an Engineering Material

Author's Note: This chapter is a semi-technical introduction to DNA for the general reader. For the technical reader, a review of the field is presented in Chapter 2.

For better or for worse, DNA biology and technology possess more lay recognition than most other sciences. Humanity's ancient and general fascination with the heredity of traits almost guaranteed that genetics would blossom as a center-stage scientific discipline even before the structure of DNA was fully unraveled. In recent times, DNA has been popularized by films such as *Jurassic Park*, and nowadays most public high school curricula teach of DNA being the "master molecule of the cell."

Given the scope and promise of DNA biotechnology, it should come as no surprise that technical improvements are improving at a rate as fast as or faster than improvements in the other great technology of our lifetime–silicon transistors. In a rough equivalent of Moore's Law, the prices of DNA synthesis and sequencing are dropping exponentially in time (Fig. 1-1).

What does come as a surprise to most who hear about my research for the first time, is the fact that I and others in my field are using DNA quite differently than the way it is used in biology. Geneticists and microbiologists primarily use synthetic DNA as a method of granting a cell the blueprints for constructing a protein that the cell would otherwise not have, using the cellular machinery for transcription and translation to process the introduced DNA. We biomolecular engineers instead use DNA as a basic programmable

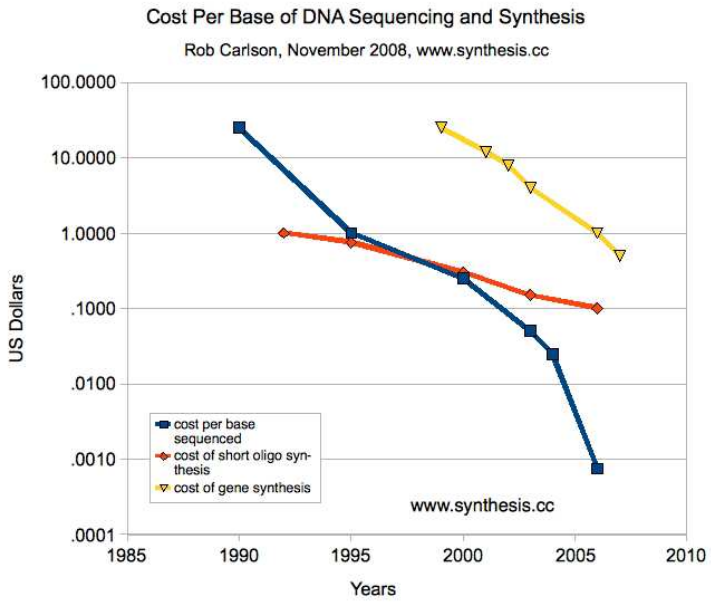


FIG. 1-1: Carlson's Law. The price of DNA oligonucleotide and gene synthesis has been dropping exponentially over the past 20 years. Image by Robert Carlson, http://www.synthesis.cc/2008/11/gene-synthesis-cost-update.html.

building block, with which we can build all sorts of useful and complex devices at the nanoscale, independent of the cell's mechanisms.

The use of naturally existing objects in ways different than their natural use is hardly a new concept; this idea dates back as far as the Stone Age, when humans used the bones of dead animals for axes and other tools. Animal bones aren't the best materials with which to make axes, but the bones were used because they were the best and most readily available material at the time. Similarly, DNA and other nucleic acids aren't the best materials with which we can make nanoscale tools, and at some point in the indefinite future, it is likely that DNA will be replaced by synthetic programmable molecules with superior properties (much like how steel has replaced bone in the axes). However, DNA is a widely and cheaply available material today, and it does have the specific and well-understood interaction properties that allow nanoscale engineering.

Velcro and Magnets

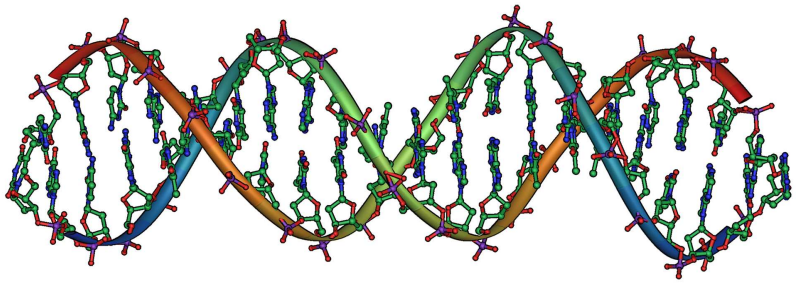


FIG. 1-2: The structure of DNA.

Whenever someone mentions the word (or, more accurately, the acronym) DNA, the image that pops to mind is usually similar to that shown in Fig. 1-2. Yes, it's true that DNA is usually in the form of a hetero-dimer molecule that adopts a double-helix shape, but thinking of DNA only as a double helix misses the point of what DNA does. This is akin to approximating a microwave as a rectangular prism, while ignoring the electronic components.

The biophysical properties of DNA that we use to perform nanoscale engineering can better be captured by analogy to Velcro and magnets (Fig. 1-3A). Velcro has two different sides, one with tiny rigid hooks, and one with flexible loops. When the two sides are brought together, the hooks latch onto the loops, and the resulting attachment is strong enough to withstand many small forces. If two hook sides or two loop sides come together, there is no resulting attachment. Similarly, magnets have north and south poles; the north poles of magnets attract the south poles of other magnets, but repel the north poles of other magnets (similarly for two south poles). We call Velcro-hooks *complementary* to Velcro-loops, and magnet-north complementary to magnet-south. In DNA, there are 4 nucleotide bases that exhibit similar complementarity: guanines (G) bind specifically to cytosine (C), and adenine (A) bind specifically to thymine (T). Thus, these 4 different components (Velcro-hook, Velcro-loop, magnet-north, magnet-south) are macroscale analogs of the 4 nucleotide bases.

Now imagine a strip of ribbon, let's say 1 inch wide and 24 inches long. We have some number of Velcro pieces and magnets, each smaller than a 1 inch square, and some double-sided tape to affix the Velcro and magnets to the ribbon. The Velcro pieces and magnets

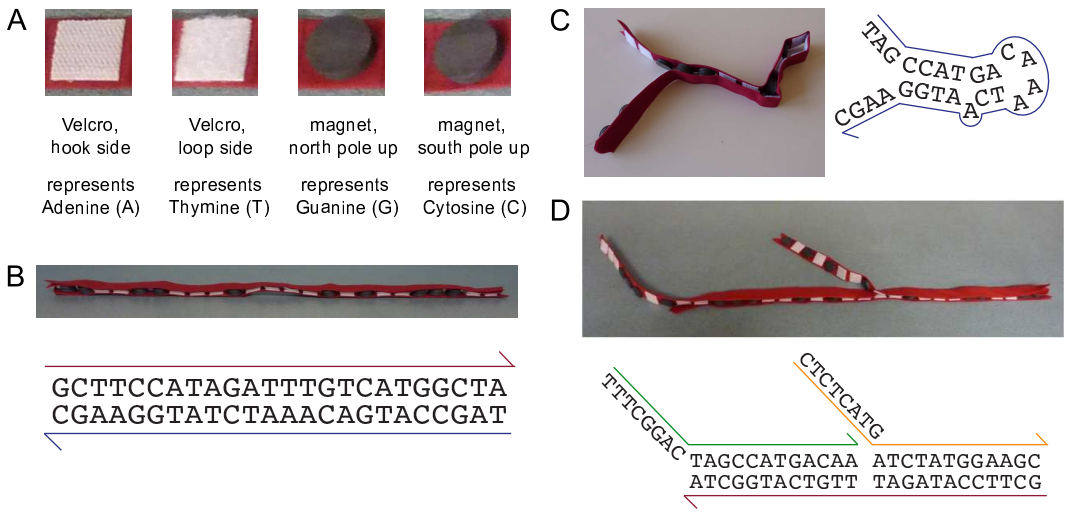


FIG. 1-3: A macroscale analog of DNA. **(A)** Two pairs of specific interactions: Velcro and magnets. Velcro has two sides, a "hook" side (H) and a "loop" side (L). Here, hooks correspond to adenines (A), and loops correspond to thymines (T). Magnets can have north pole facing up (U) or down (D). Magnets with north poles facing up correspond to guanines (G), and magnets with south poles facing up correspond to cytosines (C). A red ribbon forms the backbone of the DNA analog. The four nucleotide analogs are attached using double-sided tape (corresponding to glycoside bonds). **(B)** A functionalized ribbon matches its perfect complement at every position. Similarly, oligonucleotides bind to their perfect complements at every position. **(C)** For most sequences, a single-stranded nucleic acid can fold up on itself. **(D)** A multi-strand complex. These complexes can be separated into their component ribbons with relatively little force (as opposed to that of ripping a ribbon in half).

are placed at 1 inch intervals, so 24 different components are affixed onto the ribbon. This functionalized ribbon is a macroscale analog of a single-stranded DNA molecule that's 24 nucleotides long. The exact ordering of the squares of Velcro and magnets is known as the *sequence*.

This analog of DNA shares several important properties with DNA. First, it becomes obvious that there is an astronomical number of different sequences that we can create: for the current example, there are 24 components (nucleotides), and consequently $4^{24} \approx 2.8 \cdot 10^{14}$ different sequence possibilities. This is one advantage of working with nucleic acids: there is an **exponentially large number of different possible molecules,** leading to a **high information density**. Many other chemical polymers also have exponentially many configurations (such as halogen-substituted linear alkanes), but nucleic acids are often preferred because of (1) the ease of synthesis of aperiodic sequences, (2) the chemical distinguishability of different sequences, and (3) the predictability of functions and interactions from sequence.

Second, for each ribbon (strand) sequence, there is another sequence that is *perfectly complementary* to it: when lined up, each component (nucleotide) of the second ribbon (strand) binds to its corresponding number on the first (see Fig. 1-3D). The resulting heterodimer is an analog of canonical double-stranded DNA.

Third, for most ribbon (strand) sequences, there is substantial potential for intramolecular binding: the ribbon will get tangled because of the existence of components (nucleotides) that could bind to each other (Fig. 1-3E). A random sequence will likely cause significant intra-ribbon tangling, also known as *secondary structure*; sequences must be carefully designed in order to avoid significant secondary structure.

Finally, many different ribbons can become linked together through sections of partial complementarity (Fig. 1-3F). These multi-ribbon *complexes* are linked only by the weaker interactions between magnets and Velcro (rather than by continuous fabric), and thus can be pulled apart into individual ribbons. Similarly, multi-strand DNA complexes are metastable structures that can spontaneously "melt" at high temperatures. Complexes can be useful for the hierarchical assembly of structures, and as precursors for the controlled release of their component strands.

The Velcro and magnet-covered ribbon analog of DNA is intended to give you, the reader, a basic intuition on the thermodynamics of DNA hybridization. It is not a perfect analogy (for example, the ribbon does not inherently have a directionality as DNA does), and insights gleaned from this model are to be taken with a large helping of salt. In the next section, we shall explore the biophysics of DNA in more detail.

Before delving into the more technical aspects, however, let's first establish a set of definitions that will be used more or less throughout this thesis. DNA nanotechnology emerged at the confluence of several different fields, and it is only to be expected that various researchers would coin their own terms for common concepts before commonly accepted terminologies are established. Table 1-1 lists the names and definitions of commonly-used concepts in DNA nanotechnology, as well as synonyms used in the literature.

Biophysics and Mechanisms

Preferred term	Synonyms	Definition
strand	oligonucleotide molecule oligo	A continuous linear nucleic acid polymer covalently joined by phosphodiester bonds, typically less than 200 nucleotides.
complex	molecule tile	A non-covalently linked structure of several strands joined by Watson-Crick base pair interactions.
domain	subsequence tract section	Several continuous nucleotides in a strand that act as a unit in hybridization, branch migration, dissociation, structure, or (deoxy)ribozyme function.
toehold	sticky end	A particular type of domain that serves to colocalize nucleic acid strands and complexes. Toeholds are typically short (4 to 10 nt).

TABLE 1-1: Common terms and their synonyms and definitions.

There are three major functional differences between real nucleic acids and the ribbon analog described in the previous section (other than sheer size).

First, nucleic acids are asymmetric, directional molecules (Fig. 1-4A). Individual nucleotides are linked into strands via phosphodiester bonds that connect the 5' carbon of the (deoxy)ribose sugar to the 3' carbon of the next. Strand sequences are typically written from 5' to 3' because that is the direction in which transcription, translation, and replication occurs. Graphically, the 3' end of a nucleic acid strand is typically shown with an arrow (to denote the direction of transcription and other processes).

Two strands can only hybridize to each other if they are complementary in an anti-parallel configuration. This means that 5'− GCATTCC −3' is complementary to 5'− GGAATGC −3', and not to 5'− CGTAAGG −3'. This also means that many potential nucleic acid structures are impossible due to incorrect orientation (Fig. 1-4A). When considering potential DNA structures, it is particularly important to consider DNA orientations when evaluating the formation of structures.

Second, the thermodynamic stability afforded by Watson-Crick base pairing is primarily due not to base pairing, but rather to the π orbital stacking interactions of the aromatic rings of the nucleotide bases (Fig. 1-4B). Base stacking has been known since the 1980's to contribute to the stability of nucleic acid hybridization; Yakovchuk, Protozanova, and Frank-Kamenetskii further proposed in 2006 that the G-C and A-T base pairs actually

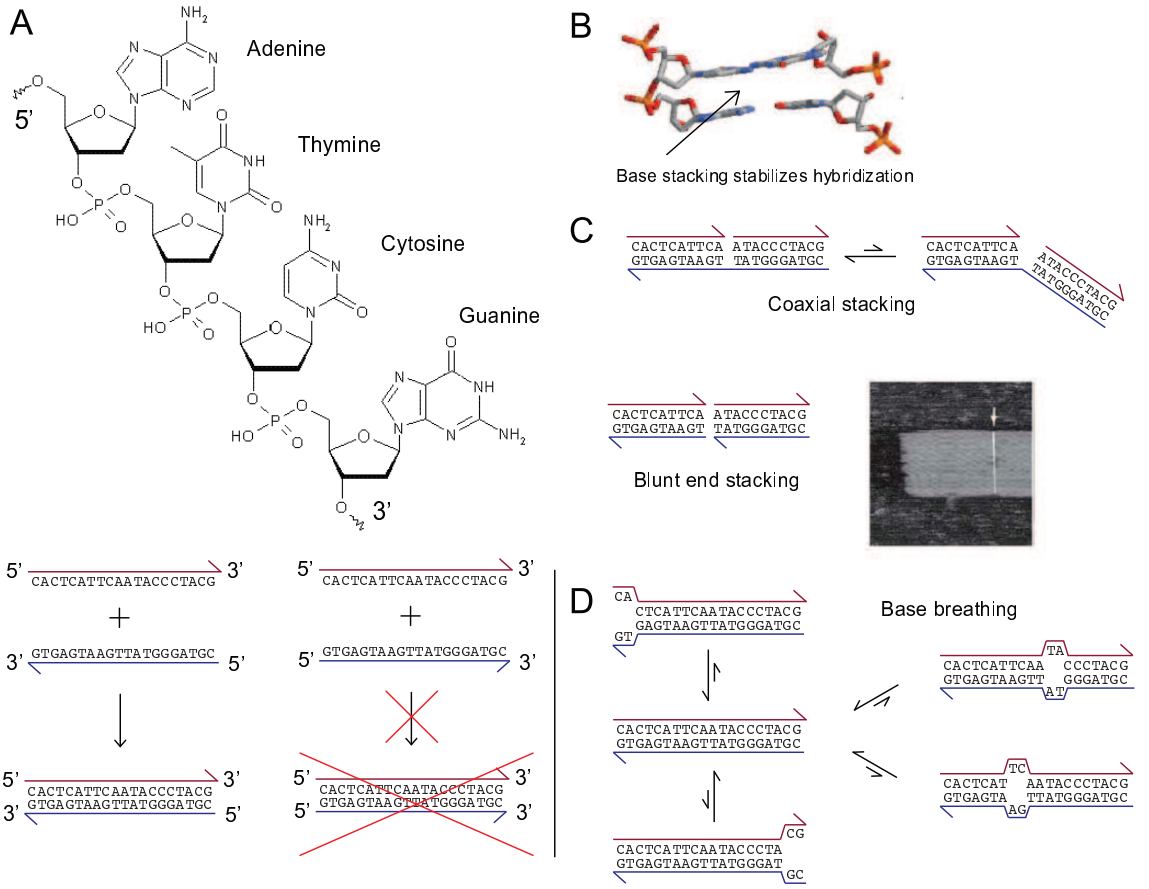


FIG. 1-4: The biophysical properties of nucleic acids. **(A)** Asymmetry and directionality of DNA. Watson-Crick base pairing requires anti-parallel orientation of DNA strands; parallel strands cannot pair. **(B)** Base stacking. The stability from nucleic acid hybridization derive mostly from the π orbital overlap from the aromatic rings on the nucleoside bases, rather than the hydrogen bonding of the complementary base pairs. Image adapted from P. Yakovchuk *et al.*, Nucleic Acids Research, 34: 564-574 (2006). **(C)** Coaxial and blunt end stacking. The shown three-stranded DNA complex possesses is composed of two disjoint helices, which adjoin at the *coaxial stack* between the two A bases. Because of stacking thermodynamics, these three-stranded structures predominantly adopt a linear rather than a kinked structure. Similarly, two different 2-stranded complexes can exhibit fleeting alignment through blunt end stacking. Usually, this cannot be observed directly, but when many different blunt end stacks are aligned, as in DNA origami, the blunt end stacks can collectively colocalize different DNA structures. Image adapted from P. W. K. Rothemund, Nature, 440: 297-301 (2006). **(D)** Base breathing. DNA bases will locally melt (breathe) fairly often because base stacking is individually weak. This breathing is transient, but has important consequences on strand displacement and branch migration kinetics.

contribute no stabilization in and of themselves. One implication of this result is that non-continuous DNA base pairs are never thermodynamically favored to form. Also, it is for this reason that DNA thermodynamics can be much more accurately predicted by nearest-neighbor models than by base composition models.

The fact that base stacking is the major contributor of DNA hybridization stability leads

to a non-obvious but important result: different DNA helices can line up due to stacking interactions. For example, a long DNA strand with 2 strands each complementary to half of it will usually appear as a rigid rod, rather than as a pair of nunchaku (Fig. 1-4C), due to *coaxial stacking*. As another example, large self-assembled DNA structures can aggregate by the ends of their helices through *blunt-end stacking*.

Finally, nucleic acid base stacks are individually weak; it is only through the simultaneous formation of many base stacks that two oligonucleotides are stably hybridized. At any particular moment, some fraction of all the DNA molecules in solutions will have some bases temporarily unbound (a.k.a. *breathing*). Base breathing is particularly prominent at the ends of helices due to the larger entropy gain.

A review of chemistry and mathematics. The next part will be a bit mathematical, so let's do a quick review of some basic chemistry math, and then ease into some field-specific math.

Let's consider a generalized chemical reaction with m reactants and n products:

$$(\text{Reactant } 1) + \cdots + (\text{Reactant } m) \underset{k_r}{\overset{k_f}{\rightleftharpoons}} (\text{Product } 1) + \cdots + (\text{Product } n)$$

This reaction has some standard free energy ΔG°, which relates to the reaction's equilibrium constant K_{eq} through the following relation:

$$\Delta G^\circ = -RT\ln(K_{eq}) \qquad \text{where } K_{eq} = \frac{\prod_{i=1}^{n}[\text{Product } i]_{eq}}{\prod_{i=1}^{m}[\text{Reactant } i]_{eq}}$$

The standard free energy of a reaction thus gives a general sense of whether the reaction is favorable or unfavorable; a negative value of ΔG° means the products are generally favored, and a positive ΔG° means the reactants are generally favored. The important exception to this rule of thumb occurs when the number of reactants and products are unequal ($m \neq n$) and concentrations are significantly different from 1 M (as often the case when dealing with nucleic acids).

To take a concrete example, consider the hybridization reaction of two complementary

DNA strands, A and $\bar{A}$:

$$A + \bar{A} \underset{k_d}{\overset{k_h}{\rightleftharpoons}} A\bar{A} \tag{1}$$

Let's suppose the ΔG° of this reaction is -4.2 kcal/mol, corresponding to $K_{eq} = 10^3$ M^{-1} at room temperature. Initially $[A] = [\bar{A}] = 1$ μM, and $[A\bar{A}] = 0$. Solving $K_{eq} = \frac{[A][\bar{A}]}{[A\bar{A}]} = 10^3$, we find that the equilibrium concentration of the product $A\bar{A}$ is only 10^{-9} M, or 0.1% of the reactant concentrations. Thus, a much more negative ΔG° is required to drive hybridization at these concentrations.

The standard free energy of a reaction can be calculated by subtracting the standard free energies of formation of the reactants from that of the products:

$$\Delta G^\circ = \sum_{i=1}^{n} \Delta G^\circ(\text{Product } i) - \sum_{i=1}^{m} \Delta G^\circ(\text{Reactant } i)$$

For the purposes of this thesis and other works dealing with only Watson-Crick interactions, we define the standard free energy of any completely unstructured oligonucleotide to be 0. From these reference points, the standard free energies of structured strands and complexes can be determined by equilibrium with unstructured strands.

The standard free energies of DNA strands and complexes (and to a lesser extent, RNA strands and complexes) have been rigorously characterized over the past 40 years. These days, one can type in the sequences of a DNA strand or complex into a Web application such as NUPACK (http://www.nupack.org/) or Mfold (http://frontend.bioinfo.rpi.edu/applications/mfold/cgi-bin/dna-form1.cgi), and receive its standard free energy within a matter of seconds, for reasonably short sequences of DNA (fewer than 200 bases). Using the standard free energies of DNA strands and complexes, we can calculate the the standard free energies of various reactions on based modifying Watson-Crick base pairing, which in turn allows us to calculated the reactions' equilibrium constants.

The equilibrium constant K_{eq} also relates the rate constants of the forward and reverse reactions: $K_{eq} = \frac{k_f}{k_r}$. Since K_{eq} can be calculated from ΔG°, and ΔG° can be calculated

from DNA sequence for reactions based on Watson-Crick interactions, the value of k_f can be calculated given k_r and vice versa.

Returning to our example hybridization reaction (1), the hybridization rate constant of two complementary strands is typically about $k_h = 10^6$ M^{-1} s^{-1} at 25 °C and 1 M NaCl. The $\Delta G^\circ = -4.2$ kcal/mol presupposed earlier (corresponding to about 5 base pairs of binding at 25 °C and 1 M NaCl) would lead to a rate constant of the dissociation reaction of $k_d = 10^3$ s^{-1}. Dissociation thus occurs on the timescale of a millisecond for 5 nt of complementarity.

In order for dissociation to be slowed to the timescale of hours ($k_d = 10^{-4}$ s^{-1}), the standard free energy of the reaction would need to be about -14 kcal/mol, corresponding to about 12 nt of complementarity. With 20 nt of complementarity, the standard free energy would be about -25 kcal/mol, and dissociation occurs at a time scale of 10^{12} s, or 30,000 years.

Branch migration and strand displacement. Consider the *strand displacement* reaction shown in Fig. 1-5A. There are more base stacks in the products ($\Delta G^\circ < 0$), and there are an equal number of reactants and products, so the products are favored at equilibrium.

$$A + B\bar{A} \underset{k_{r1}}{\overset{k_{f1}}{\rightleftharpoons}} A\bar{A} + B$$

But what about kinetics? The reactant complex $B\bar{A}$ possesses 20 nt of complementarity, which we showed in the previous section has an expected lifetime of 30,000 years. So a naive guess would be that this reaction is impractically slow, bottlenecked by the dissociation of $B\bar{A}$.

Experimentally, however, this reaction is actually quite fast: At 10 nM concentration of A and $B\bar{A}$, the reaction has a timescale of about 2 minutes. The reason for this is the base breathing phenomenon that was briefly described earlier. While the entire 20 nt complementarity is exceedingly unlikely to dissociate simultaneously, individual base pairs are rapidly breaking and reforming.

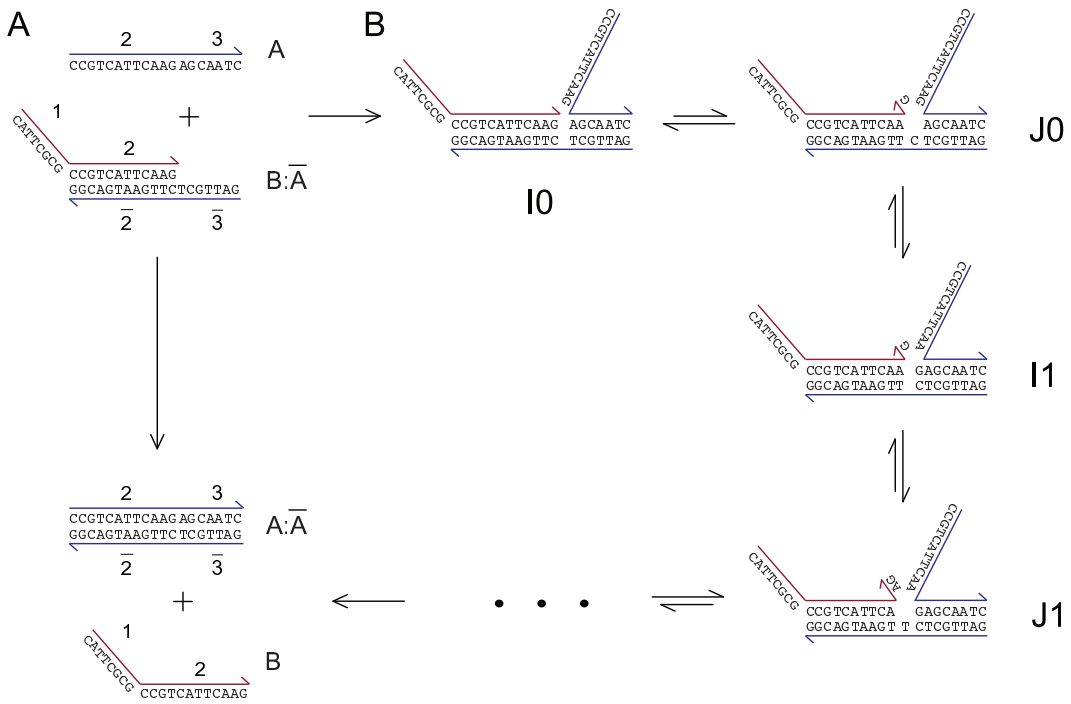


FIG. 1-5: The biophysical properties of nucleic acids. **(A)** A strand displacement reaction. **(B)** Branch migration. Branch migration occurs through a series of base breakage and reformation events. Each step of branch migration is reversible; branch migration can be accurately modeled as an unbiased random walk process.

The actual process by which this strand displacement reaction occurs is shown in Fig. 1-5B: Strand A first binds to complex $B\bar{A}$ via the 3 and $\bar{3}$ domains, forming intermediate $I0$. Next, the right-most base pair between B and $\bar{A}$ will spontaneously break with a time scale of about 100 μs, forming intermediate $J0$. The C on $\bar{A}$ now can reassociate with the G on B (to reform $I0$) or it can bind to the C on A (to form $I1$), with roughly equal probability. This base pair breaking and rearrangement process can continue with the remaining bases in the 2 domain. This series of individual base breakage and reformation steps is collectively known as *branch migration.*

Importantly, branch migration is not a directed process, but rather a random walk. From state $I1$, there's an equal probability of taking step forward to state $I2$ and of taking a step backward to return to state $I0$. Branch migration ends when all of $\bar{A}$ is bound to A, because B is then not connected to $\bar{A}$, and will diffuse away. From probability theory, the expected time to complete an unbiased random walk with distance N takes N^2x time, where x is the time needed for a single step. As mentioned previously, $x \approx 100$ μs

(at room temperature and 1 M NaCl). Consequently, branch migration over 20 nt takes approximately 40 milliseconds. For comparison, branch migration over 1000 nt would take approximately 100 seconds... still pretty fast.

Branch migration is a direct consequence of the fact that complementary nucleic acids associate by a number of individually weak interactions. In theory, branch migration can occur for any polymer exhibiting specific but weak monomer binding. Branch migration has been observed in DNA, RNA, PNA, and LNA, and is an important mechanism that enables the design of nucleic acid devices and circuits with dynamic function. Every single design presented in this thesis uses branch migration and strand displacement.

Goals of the Field

While both artists and technologists display great creativity in their respective crafts, there is a fundamental difference between the two: the latter's works must have practical applications in order to be recognized by the world. One important goal of nucleic acid bioengineering is to be able to rationally design and reprogram living organisms, through the systematic design of chemical reactions involving or templated by nucleic acids.

Nearly every aspect of cellular function is guided in some way by nucleic acids: In developmental biology, cell differentiation and body segmentation are often controlled by the relative concentrations of several protein enzymes. These proteins and other proteins that contribute the bulk of the cell's structure and function are encoded as genes in the cell's DNA. Messenger RNAs are transcribed from the cell's DNA, which in turn are translated (mostly using ribosomal RNA) into proteins. The intracellular transcription, translation, and degradation rates of nucleic acids are in turn controlled by a variety of other nucleic acids, such as microRNAs. By designing bio-orthogonal pathways that sense and influence the concentration and distribution of various nucleic acids, one can drastically change the function and fate of cells.

The short term goals of the field will be detailed in the next chapter.

Thesis Layout

The remaining chapters of this thesis describe the work on dynamic DNA devices that I've done during my graduate career at the California Institute of Technology. Rather than ordering them chronologically, I ordered them by complexity, starting with a review paper on dynamic DNA nanotechnology in Chapter 2, and ending with a whimsical proposal for the construction of a self-replicator in Chapter 10. This thesis has one Appendix chapter that describes the methods I use for nucleic acid sequence design.

The chapters of this thesis show only my more-or-less completed works in the field of DNA engineering, in which I was the major contributor and lead author. During my years as graduate student, I also authored and co-authored the following works:

---

[1] Zhang, David Y; Hariadi, Rizal; Winfree, Erik. "Programmable kinetic control of DNA self-assembly." (manuscript in preparation)

[2] Genot, Anthony J.; Zhang, David Y; Bath, Jonathan; Turberfield, Andrew J. "The remote toehold, a flexible mechanism to control hybridization kinetics." (manuscript in preparation)

[3] Seelig, Georg; Soloveichik, David; Zhang, David Y; Winfree, Erik. "Enzyme-free nucleic acid-based logic circuits." *Science* 314: 1585-1588 (2006).

[4] Schoenmyer, Tor; Zhang, David Y. "FFT-based Algorithms for the String Matching with Mismatches Problem." *J. Algorithms* 57(2): 130-139 (2005).

[5] Fujibayashi, K; Zhang, David Y; Winfree E; Murata S. "Error suppression mechanisms for DNA tile self-assembly." *Natural Computing* 8: 589-612 (2009).

[6] Zhang, David Y; Schulman, Rebecca; Cook, Matthew; Winfree, Erik. "Noisy Asynchronous Density Classification with Probabilistic One-Dimensional Block Cellular Automata." (submitted).

Chapter 2: A Survey of the Field

Nucleic acid nanotechnology can broadly be divided into three subfields that I'll call structural DNA nanotechnology, dynamic DNA nanotechnology, and interface DNA nanotechnology. Structural DNA nanotechnology is the self-assembly of nucleic acid structures with well-defined shapes, sizes, and/or patterns. Dynamic DNA nanotechnology deals with non-equilibrium systems in which DNA molecules undergo a series of conformational changes to physically or chemically modify its environment. Interface DNA nanotechnology uses nucleic acids as a tool for controlling other nanoscale materials, such as carbon nanotubes and gold nanoparticles.

In the manuscript below, Georg and I review the recent developments in DNA nanotechnology, with particular emphasis on dynamic systems. In addition to our own works, we report on many other recent exciting research results on dynamic nucleic acid devices, circuits, and motors. A majority of these dynamic designs use a common primitive, known as toehold-mediated strand displacement. Toehold-mediated strand displacement provides a reliable method of controlling DNA reaction kinetics, and is instrumental in keeping systems out of equilbrium so that the free energy of the system can be used to drive non-equilibrium dynamics.

One of our goals in writing this review paper was to present the progress of our field in terms understandable to undergraduates in the natural sciences, with the intent of recruiting them to our field. As such, we tried to write this manuscript in a way that facilitates

understanding of the presented systems: Systems are described qualitatively rather than quantitatively, and many important experimental details are omitted for the sake of understandability.

The version presented in this chapter is not a final manuscript. This work is in preparation for submission as:

Zhang, David Yu; Seelig, Georg. "Biomolecular Engineering with DNA Strand Displacement Cascades."

Kinetic control of biomolecular reaction pathways is essential to life. Similarly, control over chemical reaction kinetics is a prerequisite for building synthetic systems with complex temporal and spatial dynamics, and enables us to organize matter at the nanoscale. DNA nanotechnology achieves such kinetic control using the unique recognition properties of nucleic acids. DNA strand displacement, in particular, is a simple and robust mechanism that enables the construction of a variety of dynamically reconfigurable devices, including synthetic molecular logic circuits, catalytic amplifiers, autonomous molecular motors, and reconfigurable DNA nanostructures. Here we review these systems and explain how the strand displacement mechanism was systematically used in the rational design process.

The specificity and predictability of Watson-Crick base pairing [1] render DNA a powerful and versatile material for engineering at the nanoscale. The thermodynamics of DNA hybridization have been carefully characterized within the past decade [2–4], allowing the quantitative prediction of structures and interactions. The use of DNA as an engineering material has been additionally facilitated by exponentially decreasing cost in the preparation and purification of DNA oligonucleotides [5].

Recent years have witnessed a surge in both the number and in the scale of DNA-based nanostructures and nanodevices. In this review, we limit ourselves to describing in detail the subset of DNA nanotechnology which involve **dynamic** behavior–that is, the

constructions in which the "interesting part" is the non-equilibrium dynamics, rather than the equilibrium end-states. We primarily discuss devices that operate based on the strand displacement mechanism, because the simplicity and programmability of this mechanism facilitates the rational design of complex devices and circuits.

Strand displacement occurs when one single-stranded nucleic acid molecule hybridizes via Watson-Crick complementarity to a partially double-stranded complement. This reaction is initiated at the single-stranded toehold domains and progresses through a branch migration process. The original binding partner is released from the complement and can, in turn, trigger a downstream strand displacement event. An overview of the DNA strand displacement process is given in Box 1.

In molecular biology, strand displacement frequently denotes a process mediated by enzymes such as polymerases, but the reaction as defined above is guided by the biophysics of DNA and occurs independently of enzymes. Lee *et al.* [6] first experimentally observed DNA branch migration and strand displacement *in vitro*, and noted the isoenergetic nature of different branch migration states. Thompson *et al.* [7] and Radding *et al.* [8] modeled branch migration reactions as random walk processes, and Radding *et al.* measured the kinetics of strand displacement reactions with long toeholds (over 300 nt). Green and Tibbetts [9] realized that intrinsic secondary structure in DNA strands could frustrate branch migration, and characterized the kinetics strand displacement at elevated (65 °C) temperatures. Wetmur and coworkers [10–12] studied branch migration reactions initiated by short (3 nt and 4 nt) toeholds as a method of kinetically controlling a downstream event (in their case, ligation). Wetmur *et al.* further observed that the branch migration is sequence specific, and that the kinetics of initiating branch migration increases with stronger toehold binding thermodynamics. Reynaldo *et al.* [13] measured the kinetics of strand displacement reactions in the absence of toeholds, and observed that the kinetics of such "blunt end" strand exchange are initiated by melting of base pairs.

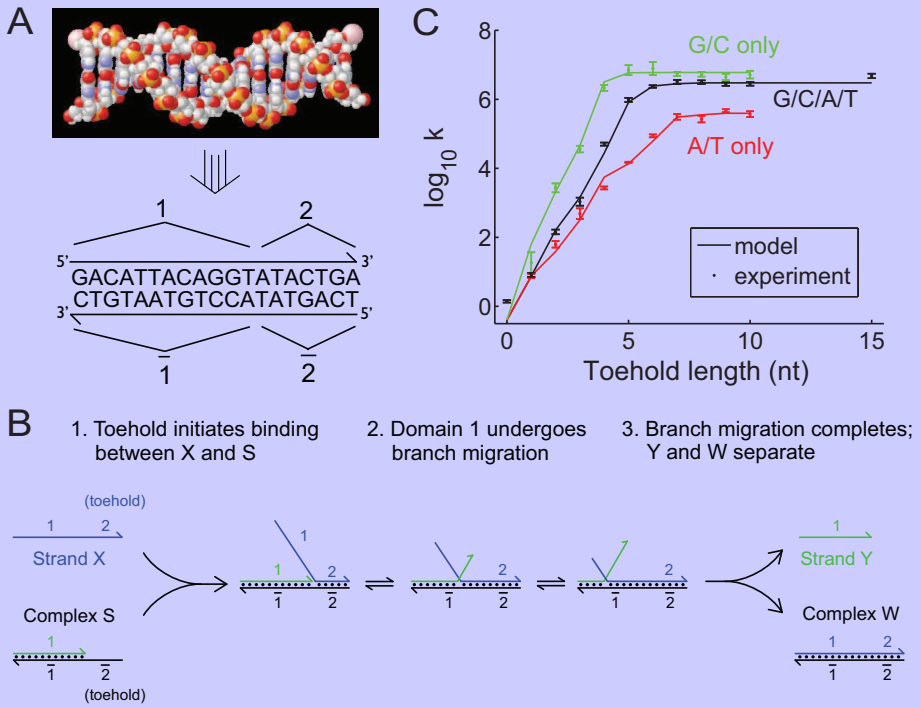
The systematic use of toeholds and strand displacement in DNA nanotechnology was pioneered by Yurke *et al.* [14], who crucially observed that the same strand of DNA can undergo multiple hybridization and strand displacement cycles. Yurke and Mills [15] observed

the kinetics of strand displacement for a variety of toehold lengths (between 0 and 6 nt), and experimentally observed rate constants differing over 4 orders of magnitude depending on the strength of the toehold binding. Zhang and Winfree [16] proposed a quantitative model relating the thermodynamics of toehold hybridization to strand displacement kinetics; furthermore, Zhang and Winfree predicted and experimentally observed saturation of kinetic speedup for toeholds longer than 7 nt (see Box 1).

Given the importance of the toehold, it seems clear that toehold inactivation could suppress strand displacement. Specific mechanisms for activating and inactivating toehold domains using strand displacement have been explored in Refs. [18, 19]. This importantly allows the cascading of strand displacement reactions: one toehold-inactivated strand could have its toehold activated via strand displacement, and subsequently participate in a downstream strand displacement reaction (see Box 2). Such cascades of strand-displacement have allowed the construction of devices and networks exhibiting more complex behaviors.

The free energy that drives strand displacement cascades is derived from the potential of forming base pairs (enthalpy gain) or releasing strands (entropy gain). The reaction is therefore limited by the amount of reactants that are supplied initially. Once the system reaches equilibrium, no more information processing or physical work can be done. Unlike with electrical circuits, strand displacement-based circuits cannot be easily recharged by hooking up to a standardized power outlet, because the reactant species used for each strand displacement reaction network will be different. Thus, strand displacement cascades are reminiscent of a domino chain reaction where toppling the first domino leads to a cascade that ends when all dominoes have fallen. This issue can be addressed using a chemostat where reactants are constantly replenished, or if mechanisms such as transcription can be used to produce more reactants.

Despite the intrinsically limited lifetime of current strand displacement cascades, it is possible to use these to perform non-trivial tasks such as amplification for oligonucleotide detection, logical evaluation for evaluating complex Boolean functions, autonomous directional movement via DNA walkers that traverse a track, and kinetically controlled self-assembly of DNA dendrimers. As an analogy, the polymerase chain reaction (PCR) is also a closed



BOX 2-1: DNA strand displacement overview.

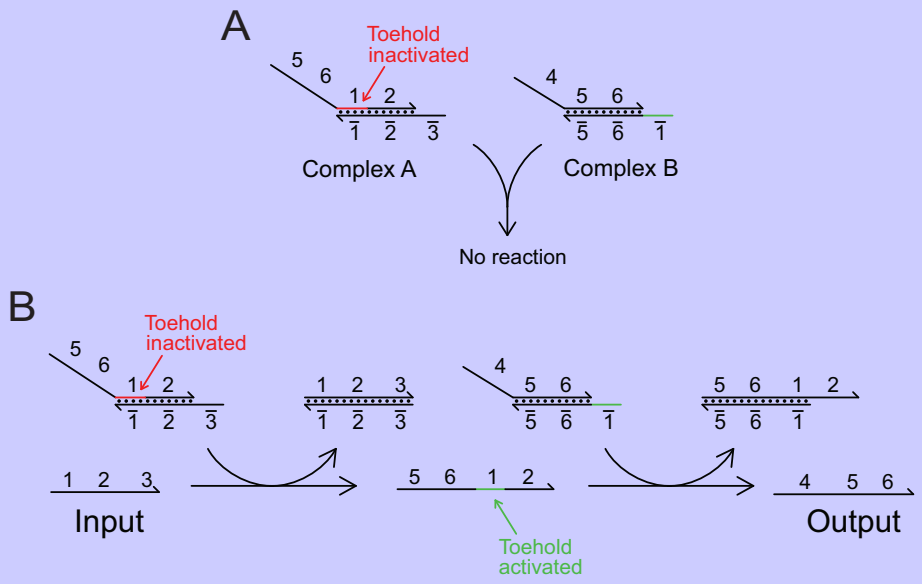
DNA is typically schematically represented as directional lines, with the hook denoting the 3' end (panel **(A)**). For many strand displacement-based designs, it is convenient to abstract contiguous DNA bases into functional DNA domains that act as a unit in hybridization, branch migration, or dissociation. Domains are represented by numbers; a barred domain denotes a domain complementary in sequence to the unbarred domain (e.g. domain $\bar{2}$ is complementary to domain 2). The sequences of the nucleotide bases are not typically shown because it is expected that strand displacement-based DNA devices will work for many if not most choices of domain sequences.

DNA strand displacement is the key primitive that has allowed the constructions presented. Panel **(B)** shows one example of this reaction. Single-stranded DNA molecule X reacts with multi-stranded DNA complex S to release strand Y and complex W. This reaction is facilitated by the "toehold" domains 2 and $\bar{2}$: The hybridization of these single-stranded toeholds colocalizes S and X, and allows the 1 domain to branch migrate. Branch migration is the random walk process in which one domain displaces another of identical sequence in binding to a complementary domain, via a series of single nucleotide dissociation and hybridization steps [8]. At the completion of branch migration, complex W is formed and strand Y is released. The use of toeholds for DNA nanotechnology was pioneered by Yurke et al. [14, 15], and characterized and modeled by Zhang and Winfree [16].

Panel **(C)** shows that the kinetics of strand displacement can be accurately modeled and predicted based on the length and sequence of the toehold domain [16]. The rate constant of the strand displacement reaction varies over 6 orders of magnitude, from 1 M^{-1} s^{-1} to $6 \cdot 10^6$ M^{-1} s^{-1}.

system, and its function depends on the presence of primer oligonucleotides, which depletes over the course of the PCR procedure.

Gain and amplification



BOX 2-2: Programming reaction cascades.

As can be seen from Box 2-1, the toehold is instrumental in controlling the kinetics of strand displacement reactions. One powerful concept that utilizes this observation is toehold inactivation, preventing strand displacement reactions by reducing the accessibility of toeholds to hybridization. Toeholds must hybridize to each other in order to serve their purpose of colocalizing DNA for branch migration; thus any mechanism that inhibits the hybridization of complementary domains will serve to inactivate the toehold. The most common method of sequestering toeholds is to simply make it double stranded.

In panel **(A)**, the 5:6:1:2 strand is prevented from reacting with complex B, because the toehold domain 1 is double-stranded. In panel **(B)**, the input strand reacts via strand displacement with complex A to release the 5:6:1:2 strand in single-stranded form. The 1 domain is now accessible, and the 5:6:1:2 strand reacts with complex B via strand displacement to release the output strand. This cascade of reactions serves as a stoichiometric nucleic acid sequence translator: the input DNA strand is completely independent in sequence of the output strand. This translator design was introduced by Seelig et al. [19] and studied in detail by Picuri et al. [17].

Biochemical amplification is integral to the function of biological reaction pathways as diverse as gene regulation, signaling, and metabolism. Engineering synthetic biochemical circuits likewise requires gain and amplification. Here we discuss how amplification and gain have been achieved through DNA strand displacement reactions. In the context of strand displacement reactions, gain is defined to be the number of nucleic acid outputs (products) released, on average, per nucleic acid input (reactant) consumed.

A general method for achieving gain is to engineer a reaction network in which the input molecule is not consumed, and can participate in multiple reactions. Then, the

input molecule can be considered as acting catalytically to release the output molecules from their inactive (partially double-stranded) states. Only base pairs are made or broken in the amplification reactions described here; this class of nucleic acid-based catalysts is quite different from (deoxy)ribozymes which generally catalyze the formation or breaking of covalent bonds (see e.g. [20] for a review of ribozymes).

Turberfield *et al.* [21] first investigated the possibility of engineering catalytic control over DNA hybridization. They showed that the hybridization of two complementary DNA strands could be slowed by constraining one or both of the strands via hybridization to shorter auxiliary strands. A specific input strand can controllably reverse this constraint, and catalytically accelerated the reaction kinetics by a factor of 30.

Several subsequent papers improved on the results of Ref. [21]. Bois *et al.* [23] and Green *et al.* [24], and Seelig *et al.* [22] reported the design of highly metastable DNA complexes that could serve as substrate for strand displacement-based catalysis, and achieved over 1000-fold catalytic speedup. This progress was in part based on the serendipitous discovery that DNA configurations reminiscent of naturally occuring "kissing hairpins" can be kinetically stable for weeks.

Dirks *et al.* [18] demonstrated kinetic control over the formation of a DNA polymer from DNA hairpin monomers. Hairpin structures reduce the accessibility of toeholds and the monomers are kinetically inhibited from spontaneously polymerizing into the more thermodynamically favorable polymer structures. The input strand triggers a chain reaction of hairpin-opening events, which leads to formation of a long double-stranded DNA polymer of up to thousands of base pairs. Venkataraman *et al.* [25] expanded on this approach to demonstrate a DNA polymer which grows by insertion of monomer units.

In most of the above systems, the catalytically produced output was of a different form than the input (double-stranded vs. single-stranded), and such catalytic strand displacement systems thus could not be easily cascaded. The catalysis system by Seelig *et al.* [22] showed the catalytic release of a single-stranded output with sequence unrelated to the input which enabled the construction of multi-layered cascades [19]. The maximal gain observed was on the order of 50; the gain of a system is here defined to be the number of

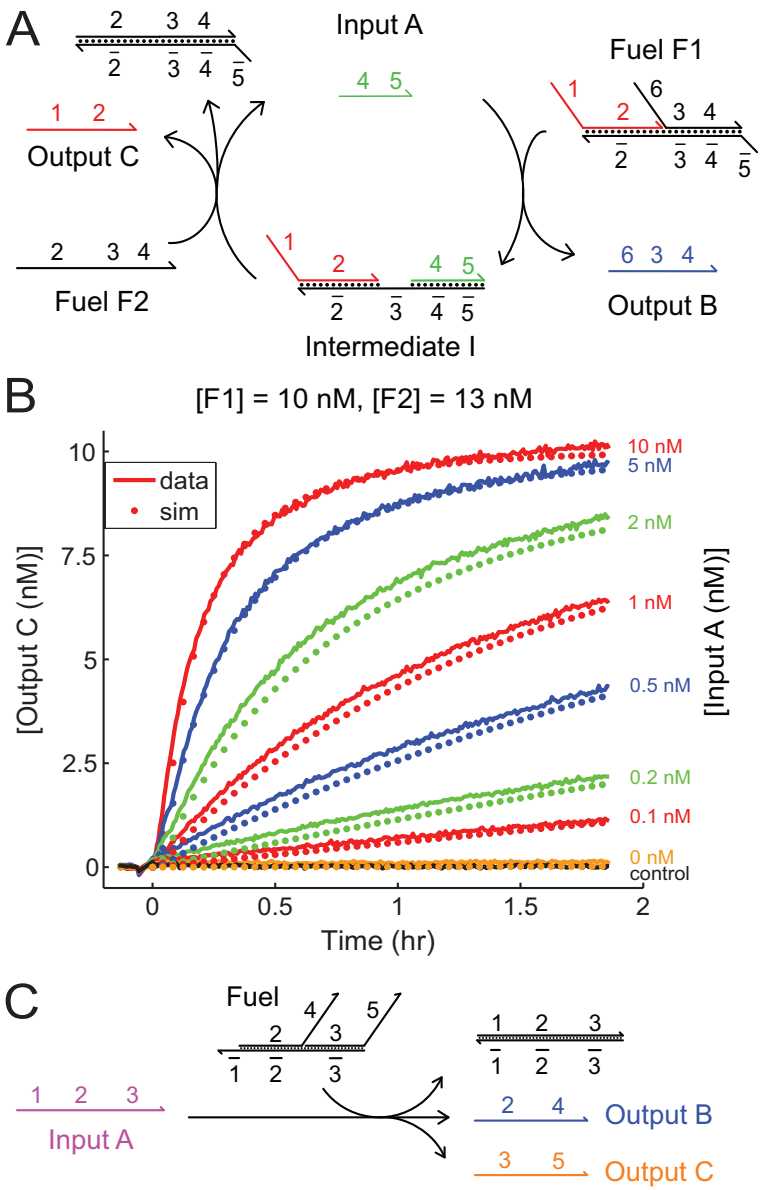


Fig. 2-1: Gain and amplification via strand displacement-based catalysis [26]. **(A)** The input strand acts catalytically to release multiple output strands of unrelated sequence. Input binds to fuel F1 to yield output B and an intermediate I. The $\bar{3}$ domain is activated in the intermediate, and allows the intermediate to react with fuel F2 to yield output C, a waste product, and the input A. The input can subsequently undergo further reaction cycles to continue releasing additional copies of the two outputs. The input can be of nearly arbitrary sequence; the sequences of fuels are designed based on that of the input. **(B)** Experimental demonstration of catalysis. The production rate of output 2 is quantitatively predicted by modeling strand displacement kinetics (dotted traces). **(C)** An alternative strand displacement-based implementation of gain that is non-catalytic. Each input molecule binds to the fuel to release two single-stranded outputs. Such reactions could potentially be cascaded to increase gain.

output strands released, on average, per input strand.

Zhang *et al.* [26] developed a catalytic system using a stable and fully rationally designed two-component substrate (see Fig. 1). Interestingly, their system was driven by the entropy

gain of additional released molecules, rather than the enthalpy gain of base pair formation. The authors experimentally demonstrated the cascading of multiple catalytic reactions, exhibiting quadratic and exponential growth kinetics. A two-stage cascaded system exhibited a gain of over 900, which can presumably be increased even further through cascading even more catalytic systems. Zhang and Winfree [27] further improved this entropy-driven catalyst system to allow dynamic allosteric modulation.

Yin *et al.* [28] demonstrated a reaction in which single-stranded DNA catalyzed the formation of multi-arm branched structures from hairpin monomers. Although the output in this case was not exactly the same form as the input, each arm of the product structures contained a single-stranded region, and could potentially act in downstream reactions. Using this technique, Yin *et al.* showed a cross-catalytic system in which two partially duplex structures catalyzed each others' formation, and exhibited exponential growth kinetics. Yin *et al.* in this paper also demonstrate other constructions such as walkers and dendrimers; these will be reviewed in subsequent sections.

Above we introduced strand-displacement catalysis as a mechanism for amplification and signal propagation. However, additional motivation for research into strand-displacement based catalysis comes from the construction of hybridization-driven DNA-based molecular motors (see "Autonomous DNA nanomachines") [21]. A biological molecular motor such as kinesin acts as a catalyst for the hydrolysis of ATP; motor movement is powered by this hydrolysis reaction. An autonomous DNA-based molecular motor similarly requires a high energy fuel source and specific coupling of movement to fuel consumption.

An alternative simple way of implementing gain is to design strand displacement reactions in which each input releases multiple outputs (Fig. 1C) [29]. The practical length constraints of oligonucleotides (based on current synthesis techniques) limit gain achieved in this way to be about 5 for a single strand displacement reaction. This limitation could potentially be overcome if multiple components with gain larger than one are cascaded.

Biomolecular Circuits

Biology uses networks of coupled chemical reactions to control cellular behavior. These

networks incorporate molecular sensors, logic elements and actuators that can detect, analyze, and respond to a variety of chemical signals. In particular, the concentrations and sequences of nucleic acids encode information in natural biological systems. Intracellular microRNA and messenger RNA expression levels, for example, provide a signature for cell type and state [30, 31]. Consequently, design of synthetic biological circuits that sense, analyze, and modulate nucleic acids can potentially be used as a powerful tool to reprogram biology. In this section we review how strand displacement cascades can be used for systematically engineering complex multi-layered reaction networks that could potentially be applied to such control problems. We will show how the reliability and predictability of the strand displacement reaction makes the construction of large multi-layered logical networks possible.

Nature provides many examples of complex biomolecular computations, ranging from neural networks in the brain to gene regulation in development. However, biological circuits are evolved, and many underlying mechanisms are not only imperfectly understood, but also difficult to systematically reproduce in a synthetic setting. Biology therefore does not provide a simple blueprint for engineering synthetic molecular circuitry. Concepts and ideas from engineering, such as the use of digital logic, provide a more practical framework for the methodical design of man-made molecular circuits.

The molecular programming work discussed here often relies on abstractions and concepts from computer science or electrical engineering. However, it is notably different in outlook and motivation from earlier demonstrations of DNA computation based on Adleman's work [32], which were geared towards solving mathematical problems difficult for electronic computers. The goal of the works reviewed here is not to compete with electronics but to implement molecular information processing similiar to that which occurs in cells.

Benenson and collaborators [33] proposed and developed a DNA and enzyme-based molecular automaton that could perform a computation where the outcome (the release or not of a mimic of an antisense drug) was dependent on the absence or presence of specific inputs (ssDNA with sequence analogous to diagnostically relevant mRNA). Stojanovic and

collaborators developed deoxyribozyme-based logic gates [34] and used these gates to implement a variety of logic circuits [35–38]. In the digital abstraction, logical values "0" and "1" are represented by low and high concentrations, respectively. While conceptually novel and elegant, these models used either enzyme or ribozyme catalysis and are therefore reliant on specific experimental conditions (such as temperature and metal ion concentrations).

In contrast, nucleic acid logic gates based only on strand displacement are more robust to different reaction conditions and thus are expected to be more easily integrated with other synthetic biological elements. Takahashi *et al.* [39] and Seelig *et al.* [19] proposed and tested designs for Boolean logic gates that use only strand displacement. Seelig *et al* further demonstrated cascaded logic circuits incorporating key design principles of digital electronics (see Fig. 2), including signal restoration which enables digital abstraction. Circuits implement a complete set of logical functions (AND, OR, and NOT) using short oligonucleotides as inputs and outputs (input sequences were chosen to be sequences of biological microRNAs). With signal restoration, a multi-component circuit was able to produce the correct output even when "noise" was introduced in the concentrations of the inputs and not all gate molecules performed perfectly (see Fig. 2C).

Frezza *et al.* [40] developed surface-bound strand-displacement based logic gates which release diffusible single-stranded signals. Notably, they use spatial separation rather than toehold inactivation to control interactions between gates. Picuri *et al.* [17] extended their approach to include toehold inactivation, and constructed a two-layer translator system for diagnostic applications.

Qian and Winfree [41] proposed a standardized method of constructing logical AND and OR gates from a basic catalytic gate motif. They further showed how arbitrary feed-forward digital logic circuits, relay contact circuits, and various analog circuits could be systematically constructed. This paper also proposed an enzyme-based method for gate assembly, which could potentially enable high-throughput production of different logic gates.

While the digital abstraction is useful for constructing reliable circuitry, chemical reactions are intrinsically analogue and their kinetics enable a much broader class of behaviors including but not limited to oscillations, chaos, and pattern formation. Soloveichik *et al.* [29]

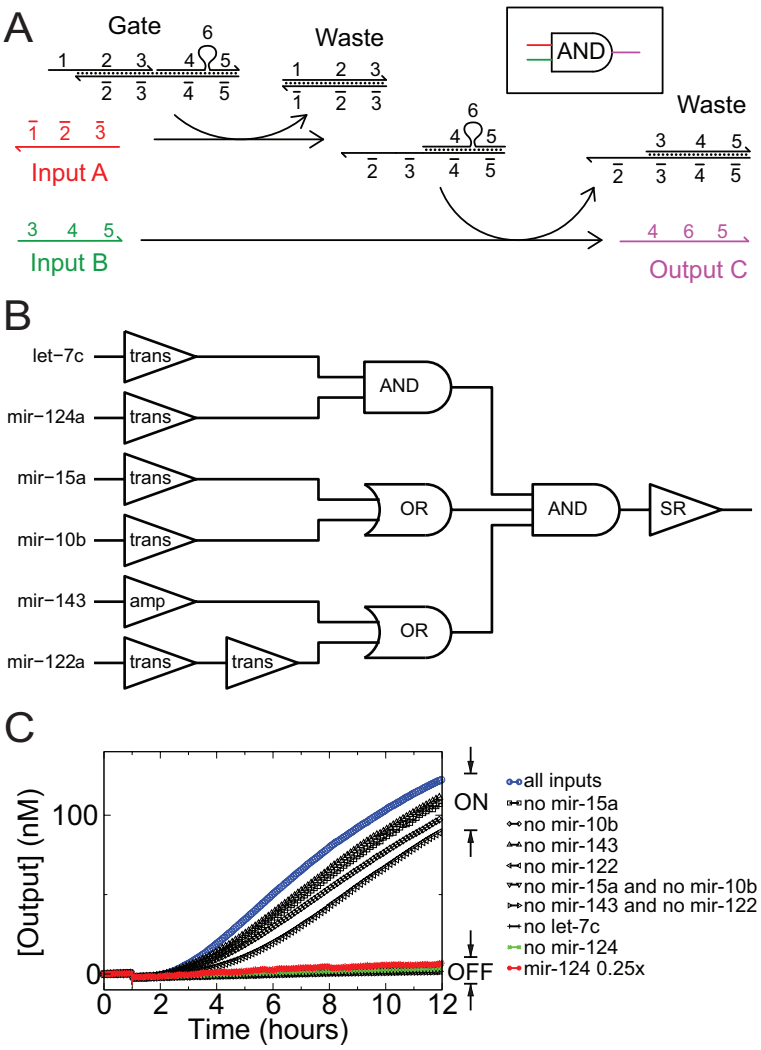


Fig. 2-2: A strand displacement circuit that performs Boolean logic [19]. **(A)** AND logic gate. Input A binds to the 3-stranded "Gate" complex at the top-left to release a waste product and activate the $\bar{3}$ toehold on the remaining 2-stranded complex. Input B then binds to the 2-stranded complex to release output C and another waste product. Output release requires two sequential strand displacement reactions: if either A or B is absent, then little C is produced. **(B)** A 12-component logical circuit. The inputs (let-7c, mir-124a, mir-15a, mir-10b, mir-143, mir-122a) denote the DNA analogs of micro RNAs, with high concentration (200 nM) denoting "ON" and low concentration (0 nM) denoting "OFF." OR logic ("OR") is implemented by using translators (Box 2-2) with different inputs but the same output. Amplification ("amp") is implemented using a kissing hairpin-based strand displacement-based catalysis reaction [22]. Signal restoration ("SR") is implemented by thresholding and amplification. **(C)** Experimental results of the 12-component circuit. The output concentration increases with time when the input concentrations are such that the logical circuit evaluates to "ON."

suggested a systematic approach for approximating arbitrary mass action kinetics through DNA strand displacement cascades. This paper treats chemical reaction kinetics as a prescriptive "programming language" and suggests an automated process for implementing a system of coupled chemical reactions with DNA-based chemistry.

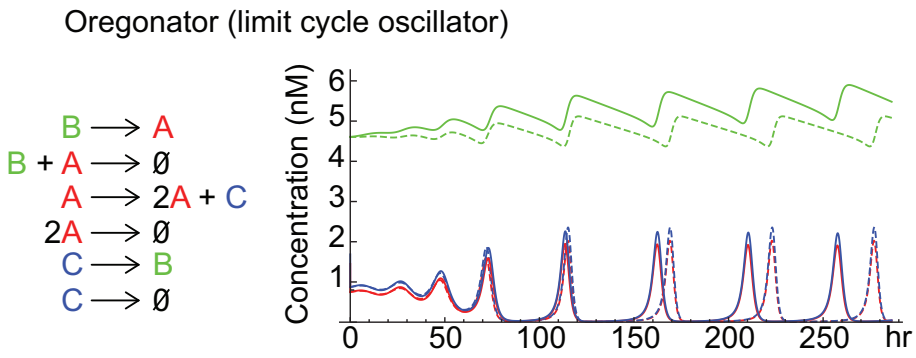


Fig. 2-3: Arbitrary chemical reaction kinetics with DNA strand displacement [29]. Simulations of the "Oregonator" reaction network, which produces limit cycle oscillations. The reactions on the left are implemented using DNA strand displacement reactions similar to those shown in Fig. 1C and Fig. 2A. The graph on the right show the simulations of both the ideal reactions (dashed traces) and the DNA implementations (solid traces). The colors of the traces match that of the species they represent (e.g. red trace denotes [*A*]).

Cardelli and various collaborators developed a stochastic model of molecular computation that is similar to the computer science concept of process algebras, which is used to coordinate actions of multiple independent agents [42]. Their formalism allows hierarchical abstraction of strand displacement reactions into functional modules, which can be used to practically "program" reaction networks to yield complex behaviors [43]. They also propose several interesting strand displacement-based reaction designs, using nicked double-stranded DNA [44], which could potentially be prepared in a high-throughput manner using nicking enzymes.

Autonomous DNA nanomachines

Protein-based molecular motors couple the release of chemical energy (e.g. hydrolysis of ATP) to mechanical work, and are essential to many cellular functions. Prominent examples include cytoskeletal motors such as kinesin walking on microtubules, rotary motors such as bacterial flagella, and polymerization motors such as actin. Synthetic molecular motors could potentially play analogous roles in choreographing molecular processes. Nucleic acids provide media for systematically engineering motors that convert chemical energy to mechanical force.

DNA hybridization and strand displacement reactions release chemical energy, but controlling this energy and coupling it to cycles of mechanical work requires ingenuity. The

hybridization of two complementary DNA strands couples energy release to mechanical work–the enthalpy gain of base pair formation causes two random-coil strands to adopt a helical conformation with significantly longer persistence length. In this example, however, the duplex formed is an inert end product that cannot be used to perform any more physical work. The two complementary strands provide both the "fuel" that drives the reaction and comprise the "motor". This lack of separation between fuel and motor in the hybridization process intrinsically means that this process, even though it converts chemical energy to physical work, cannot operate multiple cycles. In contrast, the term "motor" commonly denotes a device that performs a desired operation multiple times.

Yurke *et al.* [14] demonstrated a first example of a DNA nanomachine driven by strand displacement. Their DNA tweezers – two double-helical regions connected by a single-stranded flexible hinge – could be cycled between an open and a closed state through the repeated external addition of specific single-stranded DNA "fuel" molecules. Because sequences of the "fuel" strands for each such motor can be different, many motors can potentially operate simultaneously and independently in the same solution. This work highlights the potential of controlling DNA configurations for nanomechanical work.

Although the nanomotor presented by Yurke changes conformation at only at the nanoscale, it is possible to integrate DNA strand displacement with other materials to induce forces and physical change at the macroscale. Lin and coworkers [45] demonstrate that strand displacement can be used to dynamically modulate the stiffness of a DNA-functionalized polyacrylamide gel.

Mao *et al.* demonstrated an even earlier example of a switchable DNA nanomachine [46]. Rather than relying on strand displacement, their rotary motor takes advantage of sequences that can switch conformation between left- and right-handed helices (Z- and B-DNA, respectively) in response to changes in the ambient salt concentration. However, this motor is not sequence-specifically addressable in the way that Yurke et al.'s tweezer is. Yan and coworkers [47] used strand displacement to construct a similar rotary DNA nanomachine that does allow sequence-specific addressability.

Sherman and Seeman [48] as well as Shin and Pierce [49] used strand displacement to

implement DNA "walkers" that could be induced through the external addition of reagents to move directionally along a one-dimensional DNA track. Their designs were inspired by motor proteins such as kinesin that similarly move along one-dimensional tracks. DNA walkers could potentially act as a method for active molecular transport: in fact, Gu *et al.* [50] demonstrated that DNA walkers moving along a specific path can pick up, transport, and drop off gold nanoparticle "cargo."

The DNA nanomotors described above demonstrated that DNA can be used to control nanomechanical movement, but are not autonomous: external addition of reagents is necessary for these devices to function continuously. Researchers have continued to develop and improve non-autonomous DNA nanomotors for a variety of purposes [51–57]. These systems are interesting in their own right but synthetic molecular machines that are equally powerful as their biological counterparts need to be able to operate without human intervention.

Initial constructions of autonomous DNA walkers used enzymes or ribozymes for their operation and were driven by the formation or cleavage of covalent bonds. In particular, several groups developed DNA walkers based on enzyme or ribozyme catalysis [58–60]. With the notable exception of the walker of Ref. [60] that migrated through a 3-dimensional hydrogel matrix, these walkers typically demonstrated 2-3 steps of autonomous movement along a track. Part of the difficulty of demonstrating extended autonomous movement was the difficulty of preparing long rigid DNA tracks at the time. Since then, DNA origami technology [61] has enabled the construction of signficantly longer tracks with more complex geometry [50, 62] leading to correspondingly longer processive walks and integration of multiple different kinds of DNA nanomotors.

Recently, researchers have been able to successfully engineer mechanisms for creating DNA walkers with autonomous and directed movement using only strand displacement [28, 63–65]. Yin *et al.* [28] developed a strand displacement-based walker that moves autonomously and directionally, but has intrinsically limited processitivity (there is a roughly 50% chance that motion is terminated at every step). Omabegho demonstrated by gel electrophoresis that a walker that autonomously and processively moves 3 full steps [63]; there does not appear to be a intrinsic limitation on the number of steps this walker than

take. We illustrate the structure and function of the walker presented by Omabegho et al. in Fig. 4. The mechanism of Green *et al.* [64] is similarly processive, autonomous, and directional, and furthermore has the desirable property of track reusability.

There are three essential components for any DNA walker design: the motor that physically moves, the fuel that provides the chemical energy, and the track that prescribes the direction of motion. In many of the above designs, the track also serves as the fuel, and consequently tracks cannot be used by more than one walker without further processing. In contrast, walkers such as kinesin use a diffusible fuel (ATP), and multiple different kinesin molecules can walk simultaneously along the same microtubule. Of synthetic DNA walkers, the enzyme-based design by Yin *et al.* [59] and the strand displacement-based mechanism by Green *et al.* [64] achieve clear separation of motor, fuel, and track.

Polymerization reactions can exert physical forces and polymerization motors such as actin are common in biology [66]. Venkataraman *et al.* [25] implemented a synthetic DNA-based version of a polymerization motor that is both autonomous and processive. In a reaction pathway reminiscent of the hybridization chain reaction, monomers are sequentially inserted into a growing polymer which exerts force against a cargo (in their case, a DNA origami).

Controlling self-assembled DNA nanostructures

"Bottom-up" self-assembly methods using DNA has culminated in the development of micron-sized structures with nanometer-scale addressibility. Large scale DNA self-assembly was first demonstrated a dozen years ago using DNA double-crossover tiles [67]. Since then a number of different approaches have been used to self-assemble both 2- and 3-dimensional structures of varying sizes and complexities (see Refs. [68–70] for reviews). One approach is worth pointing out in particular, because of its scale and procedural ease [61]: Rothemund showed that the genome of the m13 single-stranded DNA virus could be used as a scaffold for constructing DNA nanostructures, with short synthetic oligonucleotides acting as splints to bring together distal regions of the m13 genome. This approach is known as "DNA origami," and has been the basis of many recent accomplishments in DNA self-assembly [56, 71–74].

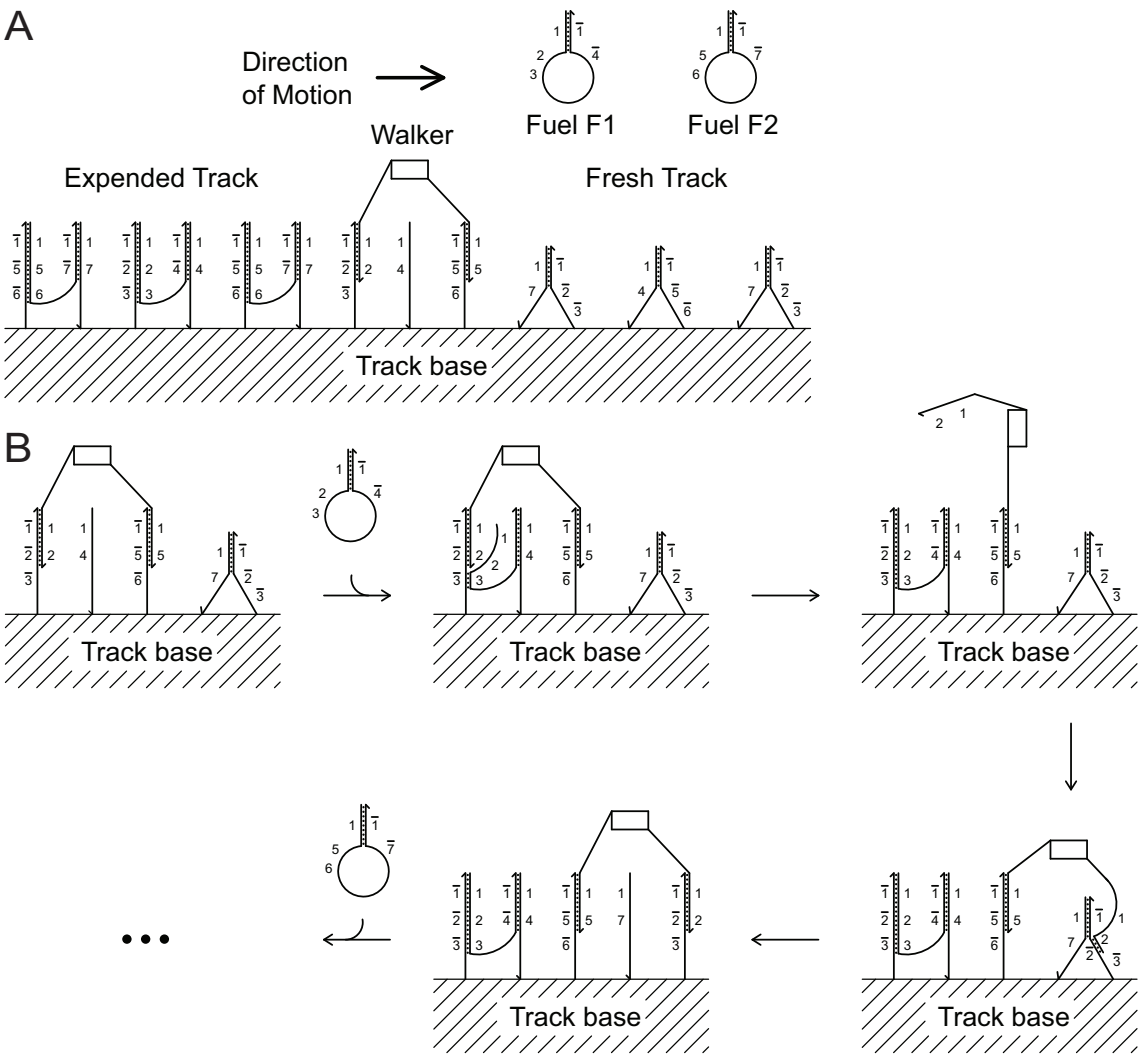


Fig. 2-4: An autonomous, processive, and directional strand displacement-based DNA walker [63]. **(A)** Fuels F1 and F2 are present in solution, and react with the track to push the walker forward. In the absence of the walker, however, the fuels do not react with the track. **(B)** Schematic of the walker taking one step. Fuel F1 displaces the hind leg of the walker through a series of strand displacement reactions. The length of the walker legs constrain the walker such that the freed hind leg cannot bind any track molecules other than that directly in front of it. The hind leg is now the leading leg, and F2 can initiate a similar reaction to drive the new hind leg forward.

In its most basic instantiations, exemplified by Winfree's original tile system or by DNA origami, DNA self assembly is driven by the equilibrium thermodynamics of hybridization. All component strands (or tiles) can be mixed at the beginning of the reaction and are favored to spontaneously assemble into the target ground state. Importantly, there is no particular order in which components react with each other. This is very different from a molecular process such as biological development or algorithmic DNA self-assembly [75] where information encoded in an emerging structure guides consecutive steps in the process.

Strand displacement cascades similarly allow one to control the timing and order of assembly processes.

Dirks and Pierce [18] pursued this approach in their hybridization chain reaction (HCR), illustrated in Fig. 5A. Yin *et al.* [28] expanded HCR to demonstrate self-assembly of branched DNA dendrimers (Fig. 5B). Combined with demonstrations of catalytic formation of multi-arm DNA structures, autonomous walkers, and exponential growth circuits, this work demonstrates the versatility of a simple hairpin motif in constructing many different types of behaviors and devices. This was achieved through integration DNA self-assembly and DNA strand displacement: each hairpin serves both as a self-assembly primitive as well as a strand displacement reactant. Lubrich *et al.* [76] demonstrated a related kinetically controlled linear polymerization process using multi-stranded monomers. Compared to traditional methods based on thermal annealing, strand displacement-based self-assembly can provide finer control over the assembly kinetics.

Strand displacement also enables dynamic reconfiguration of DNA nanostructures post-assembly. Feng *et al.* [77] report the self-assembly of a 2-dimensional DNA lattice that can dynamically change aspect ratio through the addition of single-stranded DNA that effect expansion and contraction of the monomer units. Goodman *et al.* [57] report the self-assembly of DNA tetrahedra which can be resized via toehold-mediated strand displacement of one of the edges of the tetrahedron, allowing the edge to adopt one of two different lengths depending on whether the displacing strand is present. Andersen *et al.* [56] report the construction of a 3-dimensional DNA cubic box using DNA origami technology, in which the "lid" can be opened and closed via toehold mediated strand displacement (see Fig. 6). In these works, DNA self-assembly is used to construct the "structure," while DNA strand displacement is used to implement the "moving parts." These post-assembly reconfiguration methods can be thought of as simple DNA nanomachines. The nanomachines of Yan *et al.* [47] and Lubrich *et al.* [52] are similarly used to control the state of self-assembled DNA structures.

Design tools and experimental considerations

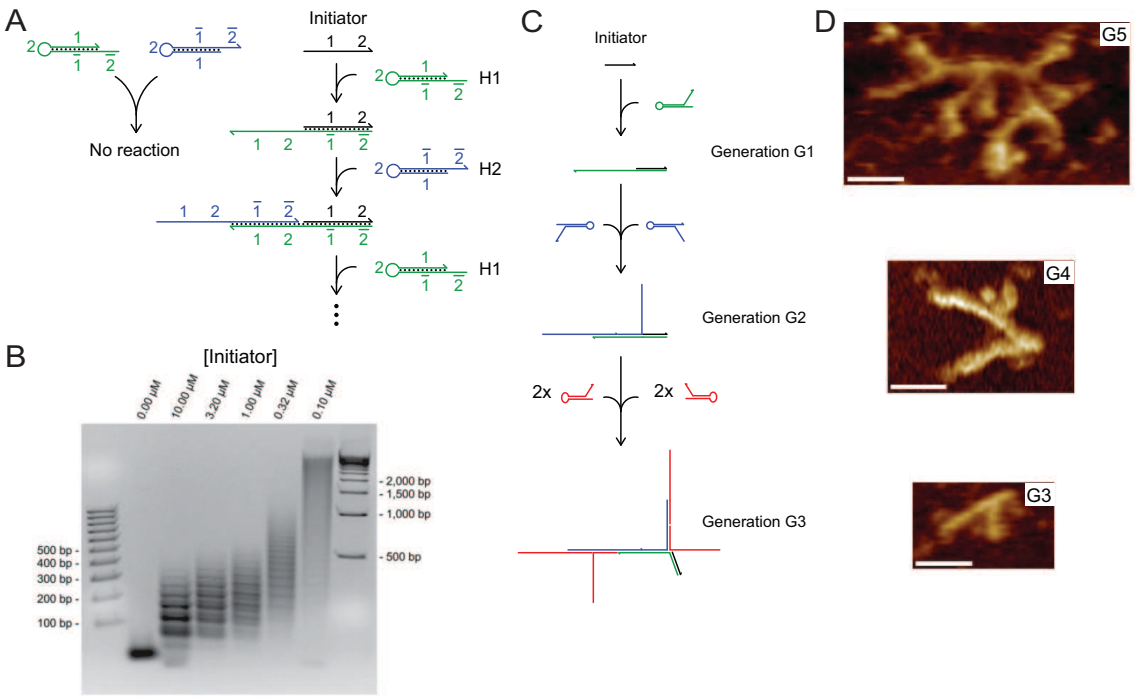


Fig. 2-5: Controlling the self-assembly of DNA nanostructures with strand displacement [18]. **(A)** Nucleated growth of linear DNA polymers, using the hybridization chain reaction (HCR). In the absence of initiator, hairpins H1 and H2 are metastable because toehold domain 2 is sequestered as a hairpin loop. The initiator reacts with hairpin H1 to activate domain 2, which then allows the polymer to react with H2, etc. **(B)** Polyacrylamide gel eletrophoresis analysis verifies that polymerization requires initiator. The size of the polymers varies inversely with the concentration of the initiator (because higher initiator concentration implies a lower hairpin to initiator concentration ratio). **(C)** Nucleated growth of branched DNA dendrimers based on the HCR [28]. The polymerization reaction begins as in the linear system, with the initiator binding to a hairpin. The extra domains on the hairpin allow the resulting linear polymer to bind to 2 additional hairpins. Each opened hairpin in turn provides 2 hairpin binding sites, and so the size of the polymer roughly doubles with every additional generation. **(D)** The right side shows the atomic force microscopy images of the 3rd, 4th, and 5th generation dendrimers.

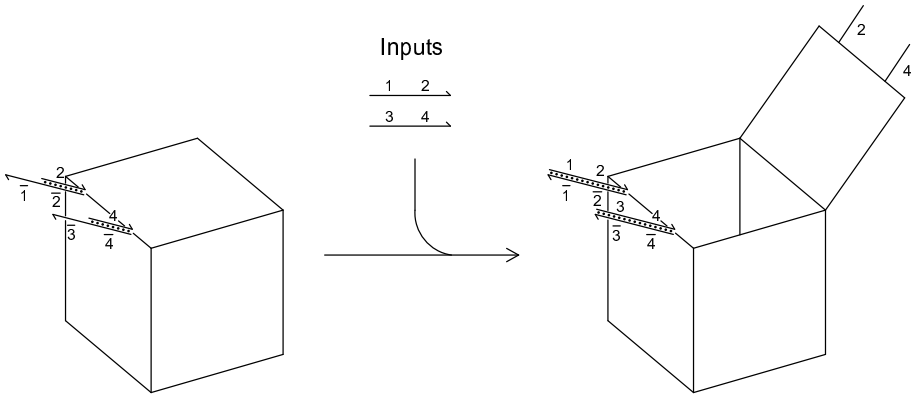


Fig. 2-6: Post-assembly reconfiguration of DNA nanostructures using strand displacement [56]. Input strands react via strand displacement with partially double-stranded complexes that keep the DNA origami box closed.

In the previous sections, we have seen various ways in which strand displacement cascades can be used to implement functional DNA devices. Here, we review design considerations and software tools that are useful for building and testing such devices. Up to this point, we have considered DNA strand displacement-based devices under idealized assumption that hybridization is perfectly specific and that no unintended hybridization occurs. Unfortunately, such sequence crosstalk is common in actual DNA systems. Sequence crosstalk hinder the kinetics of hybridization and strand displacement for all involved strands, and is generally seen to be undesirable [9, 78].

Careful sequence design is necessary to suppress sequence crosstalk, so that the kinetics of strand displacement reactions are not significantly affected [79–81]. Sequence design becomes increasingly more important but also more difficult as the complexity of DNA systems increases; this has motivated the development of automated sequence design software [79–84]. These methods generally use the thermodynamic parameters of DNA to design sequences with maximal probability of forming the desired structures and complexes at equilibrium. While not always applicable, the use of a three letter alphabet has proven to be a useful heuristic that balances maximizing sequence space with minimizing crosstalk [26, 41, 85].

Several web-based programs evaluate the thermodynamics of DNA oligonucleotides (thus testing these sequence designs) [4, 86], using nucleic acid thermodynamic parameters measured by SantaLucia and others [2]. NUPACK, developed by Dirks *et al.* [4], further analyzes the thermodynamics of multi-stranded complexes.

Another reality that DNA nanotechnologists face is the imperfection of oligonucleotide synthesis [87] – the DNA that we work with contains synthesis errors such as single-base deletions and chemical damage (e.g. deamination or depurination). Such impurities are not always completely removed even by post-synthesis purification techniques such as HPLC or PAGE. In the context of strand displacement reactions, impure oligonucleotides can lead to unintended side reactions or to slower kinetics [88]. To improve the purity of single-stranded DNA oligonucleotides, purification methods have been developed based on enzymatic proofreading [28]. For some applications it may also be possible to harness living

organisms to produce desired DNA complexes and structures; Lin *et al.* [89] show the *in vivo* cloning of basic structural motifs for DNA self assembly.

To ensure proper stoichiometry, multi-stranded complexes can be purified through native PAGE. Complexes that contain one or more defective strands may exhibit spontaneous "leak" reactions, where an output is released even without proper inputs (see e.g. [19, 26]). Such defective complexes can be selectively eliminated by pre-reaction with "trigger" strands that are metastable with correctly formed complex but react with the defective complexes [19].

The predictability of nucleic acid hybridization and strand displacement kinetics gives rise to the hope that *in silico* experiments could partially replace *in vitro* experiments in the near future. In order for this to occur, kinetic simulations of DNA reaction networks must reliably capture the behavior of DNA strands and complexes, including all potential reactions. Kinetic simulations of DNA interactions at the individual base-pairing level is one promising approach [90]. Simulations of this type should capture the dynamics of not only all designed reaction pathways, but also unintended reactions, such as from sequence crosstalk. However, such simulations are computationally expensive and may not be currently practical for large reaction networks with tens or hundreds of components.

On the other hand, even large systems can still potentially be modeled and simulated at the domain level. The DNA Strand Displacement (DSD) simulator by Phillips and Cardelli [43] is a design and simulation toolbox that automatically generates and simulates all possible strand displacement reactions given a set of strands and complexes that exist in solution. The embedded stochastic simulator plots the concentration trajectories of all initial and generated species, using experimentally measured rate constants [16].

Outlook and applications

We have shown how cascades of strand-displacement reactions can give rise to a wide range of complex behaviors from signal amplification, to chemical reaction networks and logic circuits, to molecular machines or kinetically controlled assembly. We have emphasized how the systematic use of strand displacement together with toehold sequestering is

sufficient for programming complex dynamic behaviors with Watson-Crick base pairing.

To broaden the set of chemistries that can be controlled, strand displacement circuits can be integrated with functional nucleic acids such as ribozymes, aptamers, or riboswitches [17, 18, 91]. Systems combining protein function with strand displacement further enhance the variety of behaviors that can be programmed [33, 92]. *In vitro* systems integrating multiple technologies could find applications in fields such as diagnostics or molecular imaging.

Engineered nucleic acid circuits could also be used to modulate gene expression for synthetic biological applications. While the reaction networks built so far operate in a cell-free environment, it is likely that strand-displacement based reactions networks can be adapted to work inside cells, detecting, analyzing and changing the levels of various cellular RNAs. This research will also benefit from complementary approaches currently taken in RNA synthetic biology [93–95]. Smart therapeutics applications such as those suggested in Refs [33, 36] are a particularly promising area. For this, sensors and logic circuits similar to those already available need to be integrated with molecular actuators based on antisense oligonucleotides, siRNA or ribozymes. First steps in this direction have been taken [96, 97].

In vivo operation may benefit from the use of chemically modified nucleic acids such as LNA, PNA, or 2'-O-Methyl RNA to increase nuclease stability or to modify thermodynamic properties. Such modified nucleic acids may also be of interest for cell free application as sensors or simply as an additional degree of freedom for controlling thermodynamic stability and reaction kinetics of various systems. Recent research on expanded nucleic acid alphabets [98] could likewise form the basis for bio-orthogonal embedded computation systems. The strand displacement reactions and systems described in this review are expected to function qualitatively similarly for these other nucleic acids [99].

A variety of nanoscale materials have been previously integrated with nucleic acids, taking advantage of specific Watson-Crick base pairing to direct the colocalization of other nanoscale materials. Liu and coworkers, for example, used the specific hybridization of nucleic acids to speed up the reactions of organic molecules functionalized to the complementary strands, in a process known as DNA templated synthesis [100]. Mirkin and coworkers used DNA hybridization to direct the aggregation of gold nanoparticles, which enabled a

colorimetric assay of nucleic acids [101]. Yan and coworkers showed that self-assembled DNA nanostructures can be used to direct precise spatial arrangement of proteins [74] and Le *et al.* used DNA to assemble arrays of gold nanoparticles [102]. Nucleic acid logical or catalytic circuits could allow finer control, faster kinetics, and sensitive conditional reactions for all of these processes.

Acknowledgements. DYZ is supported by the Fannie and John Hertz Foundation. GS is supported by a Career Award at the Scientific Interface from the Burroughs Wellcome Fund and an NSF CAREER award.

[1] Bloomfield, V. A., Crothers & D. M., Tinoco, I. Jr. *Nucleic Acids: Structures, Properties, and Functions* University Science Books, Sausalito, CA (2000).

[2] SantaLucia, J. & Hicks, D. The Thermodynamics of DNA Structural Motifs. *Ann. Rev. Biochem.* **33,** 415-440 (2004).

[3] Owczarzy, R., Moreira, B. G., You, Y., Behlke, M. A., Walder, J. A. Predicting stability of DNA duplexes in solutions containing magnesium and monovalent cations. *Biochemistry* **47,** 5336-5353 (2008).

[4] Dirks, R. M., Bois, J. S., Schaeffer, J. M., Winfree, E. & Pierce, N. A. Thermodynamic Analysis of Interacting Nucleic Acid Strands. *SIAM Review* **49,** 65-88 (2007).

[5] Carlson, R. The changing economics of DNA synthesis. *Nat. Biotech.* **27,** 1091-1094 (2009).

[6] Lee, C. S., Davis, R. W. & Davidson, N. A physical study by electron microscopy of terminally repetitious, circularly permuted DNA from coliphage particles of Escherichia-coli 15. *J. Mol. Biol.* **48,** 1-22 (1970).

[7] Thompson, B. J., Camien, M. N. & Warner, R. C. Kinetics of branch migration in double-stranded DNA. *Proc. Nat. Acad. Sci.* **73,** 2299-2303 (1976).

[8] Radding, C. M., Beattie, K. L., Holloman W. K. & Wiegand, R. C. Uptake of homologous single-stranded fragments by superhelical DNA. *J. Mol. Biol.* **116,** 859-839 (1977).

[9] Green, C. & Tibbetts, C. Reassociation rate limited displacement of DNA strands by branch migration. *Nucleic Acids Research* **9,** 1905 (1981).

[10] Quartin, R. S., Plewinska, M. & Wetmur, J. G. Branch migration mediated DNA labeling and cloning. *Biochemistry* **28,** 8676-8682 (1989).

[11] Weinstock, P. H. & Wetmur, J. G. Branch capture reactions - effect of recipient structure. *Nucleic Acids Research* **18,** 4207-4213 (1990).

[12] Wong, D. M., Weinstock, P. H. & Wetmur, J. G. Branch capture reactions: displacers derived from asymmetric PCR. *Nucleic Acids Research* **19,** 2251-2259 (1991).

[13] Reynaldo, L. P., Vologodskii, A. V., Neri, B. P., Lyamichev, V. I. The kinetics of oligonucleotide replacements, *J. Mol. Biol.* **297,** 511-520 (2000).

[14] Yurke, B., Turberfield, A. J., Mills, A. P., Simmel, F. C. & Neumann, J. L. A DNA-fuelled molecular machine made of DNA. *Nature* **406,** 605-608 (2000).

[15] Yurke, B. & Mills, A. P. Using DNA to Power Nanostructures, *Genetic Programming and Evolvable Machines* **4,** 111-122 (2003).

[16] Zhang, D. Y. & Winfree, E. Control of DNA Strand Displacement Kinetics Using Toehold Exchange. *J. Am. Chem. Soc.* **131,** 17303-17314 (2009).

[17] Picuri, J. M., Frezza, B. M. & Ghadiri, M. R. Universal translators for nucleic acid diagnosis. *J. Am. Chem. Soc.* **131,** 9368-9377 (2009).

[18] Dirks, R. M. & Pierce, N. A. Triggered Amplification by Hybridization Chain Reaction. *Proc. Nat. Acad. Sci.* **101,** 15275-1278 (2004).

[19] Seelig, G., Soloveichik, D., Zhang, D. Y. & Winfree, E. Enzyme-free nucleic acid logic circuits. *Science* **314,** 1585-1588 (2006).

[20] Serganov, A. & Patel, D. J. Ribozymes, riboswitches, and beyond: regulation of gene expression without proteins. *Nat. Rev. Genet.* **8,** 776-790 (2007).

[21] Turberfield, A. J., Mitchell, J. C., Yurke, B., Mills, A. P., Blakey, M. I. & Simmel, F. C. DNA Fuel for Free-running Nanomachines. *Phys. Rev. Lett.* **90,** 118102 (2003).

[22] Seelig, G., Yurke, B. & Winfree, E. Catalyzed Relaxation of a Metastable DNA fuel. *J. Am. Chem. Soc.* **128,** 12211-12220 (2006).

[23] Bois, J. S., Venkataraman, S., Choi, H. M. T., Spakowitz, A. J., Wang, Z. G. & Pierce, N. A. Topological Constraints in Nucleic Acid Hybridization Kinetics. *Nuc. Acid Res.* **33,** 4090-4095 (2005).

[24] Green, S. J, Lubrich, D. & Turberfield, A. J. DNA hairpins: fuel for autonomous DNA devices. *Biophysical Journal* **91,** 2966-2975 (2006).

[25] Venkataraman, S., Dirks, R. M., Rothemund, P. W. K., Winfree, E. & Pierce, N. An autonomous polymerization motor powered by DNA hybridization. *Nature Nanotech.* **2,** 490-494 (2007).

[26] Zhang, D. Y., Turberfield, A. J., Yurke, B. & Winfree, E. Engineering entropy-driven reactions and networks catalyzed by DNA. *Science* **318,** 1121-1125 (2007).

[27] Zhang, D. Y. & Winfree, E. Dynamic allosteric control of noncovalent DNA catalysis reactions. *J. Am. Chem. Soc.* **130,** 13921-13926 (2008).

[28] Yin, P., Choi, H. M. T., Calvert, C. R. & Pierce, N. A. Programming biomolecular self-assembly pathways. *Nature* **451,** 318-322 (2008).

[29] Soloveichik, D., Seelig, G. & Winfree, E. DNA as a universal substrate for chemical kinetics. *Proc. Nat. Acad. Sci.* (in press, 2010).

[30] Bartel, D. P. MicroRNAs: target recognition and regulatory functions. *Cell* **136,** 215-233 (2009).

[31] Lu, J. *et al.* MicroRNA expression profiles classify human cancers. *Nature* **435,** 834-838 (2005).

[32] Adleman, L. M. Molecular computation of solutions to combinatorial problems. *Science,* **266,** 1021-1024 (1994).

[33] Benenson, Y., Gil, B., Ben-Dor, U. Adar, R. & Shapiro, E. An autonomous molecular computer for logical control of gene expression. *Nature* **429,** 423-429 (2004).

[34] Stojanovic, M. N., Mitchell, T. E. & Stefanovic D. Deoxyribozyme-based logic gates. *J. Am. Chem. Soc.* **124,** 3555-3561 (2002).

[35] Stojanovic, M. N. & Stefanovic D. A deoxyribozyme-based molecular automaton. *Nat. Biotechnol.* **21,** 1069-1074 (2003).

[36] Stojanovic, M. N., Semova, S., Kolpashchikov, D., Macdonald, J., Morgan, C. & Stefanovic, D. Deoxyribozyme-based ligase logic gates and their initial circuits. *J. Am. Chem. Soc.* **127,** 6914-6915 (2005).

[37] Lederman, H., Macdonald, J., Stephanovic, D. & Stojanovic, M. N. Deoxyribozyme-based three-input logic gates and construction of a molecular full adder. *Biochemistry* **45,** 1194-1199 (2006).

[38] Yashin, R., Rudchenko, S. & Stojanovic, M. N. Networking particles over distance using oligonucleotide-based devices. *J. Am. Chem. Soc.* **129,** 15581-15584 (2007).

[39] Takahashi, K., Yaegashi, S., Kameda, A. & Hagiya, M. Chain Reaction Systems Based on Loops Dissociation of DNA. *Lecture Notes of Computer Science* **3892,** 347-358 (2006).

[40] Frezza, B. M., Cockroft, S. L. & Ghadiri, M. R. Modular multi-level circuits from immobilized DNA-based logic gates. *J. Am. Chem. Soc.* **129,** 14875-14879 (2007).

[41] Qian, L. & Winfree, E. A simple DNA gate motif for synthesizing large-scale circuits. *Lecture Notes of Computer Science* **5347,** 70-89 (2009).

[42] Cardelli, L. Strand algebras for DNA computing. *Lecture Notes of Computer Science* **5877,** 12-24 (2008).

[43] Phillips, A. & Cardelli, L. A programming language for composable DNA circuits. *Journal of the Royal Society Interface* **6,** S419-S436 (2009).

[44] Cardelli, L. "Two-domain DNA strand displacement," in Cooper, S. B., Kashefi, E. & Panangaden, P. *Developments in Computational Models*. EPTCS **25,** 33-47 (2010).

[45] Lin, D. C., Yurke, B. & Langrana, N. A. Mechanical properties of a reversible, DNA-crosslinked polyacrylamide hydrogel. *J. Biomech. Eng.* **126,** 104-110 (2004).

[46] Mao, C., Sun, W., Shen, Z. & Seeman, N. C. A nanomechanical device based on the B-Z transition of DNA. *Nature* **397,** 144-146 (1999).

[47] Yan, H., Zhang, X. P., Shen, Z. Y & Seeman, N. C. A robust DNA mechanical device controlled by hybridization topology. *Nature* **415,** 62-65 (2002).

[48] Sherman, W. B. & Seeman, N. C. A precisely controlled DNA biped walking device. *Nano Lett.* **4,**

1203-1207 (2004).

[49] Shin, J S. & Pierce, N. A. A synthetic DNA walker for molecular transport. *J. Am. Chem. Soc.* **126,** 10834-10835 (2004).

[50] Gu, H., Chao. J., Xiao, S. & Seeman, N. C. A proximity-based programmable DNA nanoscale assembly line. *Nature* **465,** 202-205 (2010).

[51] Chen, Y., Lee, S. H. & Mao, C. A DNA nanomachine based on a duplex-triplex transition. *Angew. Chem. Int. Ed.* **43,** 5335-5338 (2004).

[52] Lubrich, D., Lin, J. & Yan, J. A contractile DNA machine. *Angew. Chemie Int. Ed.* **47,** 7026-7028 (2008).

[53] Liu, H. & Liu, D. S. DNA nanomachines and their functional evolution. *Chem. Commun.* **19,** 2625-2636 (2009).

[54] Bath, J & Turberfield, A. J. DNA nanomachines. *Nature Nanotech.* **2,** 275-284 (2007).

[55] Klapper, Y., Sinha, N., Ng, T. W. S. & Lubrich, D. A rotational DNA nanomotor driven by an externally controlled electric field. *Small* **6,** 44-47 (2010).

[56] Andersen, E. S. *et al.* Self-assembly of a nanoscale DNA box with a controllable lid. *Nature* **459,** 73-76 (2009).

[57] Goodman, R. P., Heilemann, M., Doose, S., Erben, C. M., Kapanidis, A. N & Turberfield, A. J. Reconfigurable, braced, three-dimensional DNA nanostructures. *Nature Nanotech.* **3,** 93-96 (2008).

[58] Tian, Y., He, Y. Chen, Y., Yin, P. & Mao, C. D. Molecular devices - a DNAzyme that walks processively and autonomously along a one-dimensional track. *Angew. Chem. Int. Ed.* **44,** 4355-4358 (2005).

[59] Yin, P., Yan, H., Daniell, X. G., Turberfield, A. J. & Reif, J. H. A unidirectional DNA walker that moves autonomously along a track. *Angew. Chem. Int. Ed.* **43,** 4906-4911 (2004).

[60] Pei, R., Taylor, S. K., Stefanovic, D., Rudchenko, S., Mitchell, T. E. & Stojanovic, M. N. Behavior of Polycatalytic Assemblies in a Substrate-Displaying Matrix. *J. Am. Chem. Soc.* **128,** 12693-12699 (2006).

[61] Rothemund, P. Folding DNA to create nanoscale shapes and patterns. *Nature* **440,** 297-302 (2006).

[62] Lund, K. *et al.* Molecular robots guided by prescriptive landscapes. *Nature* **465,** 206-210 (2010).

[63] Omabegho, T., Sha, R. & Seeman, N. C. A Bipedal DNA Brownian Motor with Coordinated Legs. *Science* **324,** 67-71 (2009).

[64] Green, S., Bath, J. & Turberfield, A. J. Coordinated Chemomechanical Cycles: A Mechanism for Autonomous Molecular Motion. *Phys. Rev. Lett. 101,* 238101 (2008).

[65] Bath, J., Green, S. J., Allen, K. E. & Turberfield, A. J. Mechanism for a Directional, Processive, and Reversible DNA Motor. *Small* **5,** 1513-1516 (2009).

[66] Theriot, J. A. The polymerization motor. *Traffic* **1,** 19-28 (2000).

[67] Winfree, E., Liu, F., Wenzler, L. A. & Seeman, N. C. Design and self-assembly of two-dimensional

DNA crystals. *Nature* **394,** 539-544 (1998).

[68] Seeman, N. C. An overview of structural DNA nanotechnology. *Mol. Biotech.* **37,** 246-257 (2007).

[69] Aldaye, F. A., Palmer, A. L. & Sleiman, H. F. Assembling materials with DNA as the guide. *Science* **321,** 1795-1799 (2008).

[70] Shih, W. M. & Lin, C. Knitting complex weaves with DNA origami. *Curr. Opin. Struct. Biol.* **131,** doi 10.1016/j.sbi.2010.03.009 (2010).

[71] Douglas, S. M., Dietz, H., Liedl, T., Hogberg, B., Graf, F. & Shih, W. M. Self-assembly of DNA into nanoscale three-dimensional shapes. *Nature* **459,** 414-418 (2009).

[72] Dietz, H., Douglas, S. M. & Shih W. M. Folding DNA into twisted and curved nanoscale shapes. *Science* **325,** 725-730 (2009).

[73] Maune, H. T., Han, S., Barish, R. D., Bockrath, M., Goddard, W. A., Rothemund, P. W. K. & Winfree, E. Self-assembly of carbon nanotubes into two-dimensional geometries using DNA origami templates. *Nature Nanotech.* **5,** 61-66 (2010).

[74] Rinker, S., Ke, Y., Liu, Y., Chhabra, R. & Yan, H. Self-assembled DNA nanostructures for distance-dependent multivalent ligand-protein binding. *Nature Nanotech.* **3,** 418-422 (2008).

[75] Rothemund, P. W. K., Papadakis, N. & Winfree, E. Algorithmic self-assembly of DNA Sierpinski triangles. *Plos Biol.* **2,** 2041-2053 (2004).

[76] Lubrich, D., Green, S. J. & Turberfield, A. J. Kinetically controlled self-assembly of DNA oligomers. *J. Am. Chem. Soc.* **131,** 2242-2243 (2009).

[77] Feng, L. P., Park, S. H., Reif, J. H. & Yan, H. A two-state DNA lattice switched by DNA nanoactuator. *Angew. Chem. Int. Ed.* **42,** 4342-4346 (2003).

[78] Gao, Y., Wolf, L. K. & Georgiadis, R. M. Secondary structure effects on DNA hybridization kinetics: a solution versus surface comparison. *Nuc. Acids Res.* **34,** 3370-3377 (2006).

[79] Seeman, N. C. De novo design of sequences for nucleic acid structural engineering. *J. Biomol. Struct. Dyn.* **8,** 573-581 (1990).

[80] Dirks, R. M., Lin, M., Winfree, E. & Pierce, N. A. Paradigms for computational nucleic acid design. *Nuc. Acids Res.* **32,** 1392-1403 (2004).

[81] Tulpan, D., Andronescu, M., Chang, S. B., Shortreed, M. R., Condon, A., Hoos, H. H. & Smith, L. M. Thermodynamically based DNA strand design. *Nuc. Acids Res.* **33,** 4951-4964 (2005).

[82] Seifferf, J. & Huhle, A. A full-automatic sequence design algorithm for branched DNA structures. *J. Biomol. Struct. Dyn.* **25,** 453 (2008).

[83] Tanaka, F. Kameda, A. Yamamoto, M. & Ohuchi, A. Design of nucleic acid sequences for DNA computing based on a thermodynamic approach. *Nuc. Acids Res.* **33,** 903 (2005).

[84] Sager, J. & Stefanovic, D. Designing nucleotide sequences for computation: a survey of constraints. *Lecture Notes of Computer Science* **3892,** 275-289 (2006).

[85] Mir, K. U. A restricted genetic alphabet for DNA computing. *DNA based computers II. DIMACS* **44,**

243-246 (1998).

[86] Zuker, M. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.* **31,** 3406-15, (2003).

[87] Temsamani, J., Kubert, M. & Agrawal, S. Sequence identity of the n-1 product of a synthetic oligonucleotide. *Nuc. Acids Res.* **23,** 1841-1844 (1995).

[88] Zhang, D. Y. & Winfree, E. Robustness and modularity properties of a non-covalent DNA catalytic reaction. *Nuc. Acid Res.* (published online, doi:10.1093/nar/gkq088, 2010).

[89] Lin, C., Rinker, S., Wang, X., Liu, Y., Seeman, N. C. & Yan, H. In vivo cloning of artificial DNA nanostructures. *Proc. Nat. Acad. Sci.* **105,** 17626-17631 (2008).

[90] Flamm, C., Fontana, W., Hofacker, I. L & Schuster, P. RNA folding at elementary step resolution. *RNA* **6,** 325-338 (2000).

[91] Penchovsky R. & Breaker R. R. Computational design and experimental validation of oligonucleotide-sensing allosteric ribozymes. *Nature Biotech.* **23,** 1424-1433 (2005).

[92] Kim, J., White, K. S. & Winfree, E. Construction of an in vitro bistable circuit from synthetic transcriptional switches. *Molecular Systems Biology* **2,** 68-79 (2006).

[93] Beisel, C. L. & Smolke, C. D. Design principles for riboswitch function. *PLoS Computational Biology* **5,** e1000363 (2009).

[94] Isaacs, F. J., Dwyer, D. J., Ding, C. M., Pervouchine, D. D., Cantor, C. R. & Collins, J. J. Engineered riboregulators enable post-transcriptional control of gene expression. *Nature Biotech.* **22,** 841-847 (2004).

[95] Isaacs, F. J., Dwyer, D. J. & Collins, J. J. RNA synthetic biology. *Nature Biotech.* **24,** 545-554 (2006).

[96] Xie, Z., Liu, S. J., Bleris L. & Benenson, Y. Logic integration of mRNA signals by an RNAi-based molecular computer. *Nuc. Acids Res.* **38,** 2692-2701 (2010).

[97] Win, M. N. & Smolke, C. D. Higher-order cellular information processing with synthetic RNA devices. *Science* **322,** 456-460 (2008).

[98] Krueger, A. T. & Kool, E. T. Redesigning the Architecture of the Base Pair: Toward Biochemical and Biological Function of New Genetic Sets. *Chem Biol.* **16,** 242-248 (2009).

[99] He, G., Rapireddy, S., Bahal, R., Sahu, B. & Ly, D. H. Strand invasion of extended, mixed-sequence B-DNA by gamma PNAs. *J. Am. Chem. Soc.* **131,** 12088-12090 (2009).

[100] Gartner, Z. J., Tse, B. N., Grubina, R., Doyon, J. B., Snyder, T. M & Liu, D. R. DNA-Templated Organic Synthesis and Selection of a Library of Macrocycles. *Science* **305,** 1601-1605 (2004).

[101] Elghanian, R.; Storhoff, J. J.; Mucic, R. C.; Letsinger, R. L.; Mirkin, C. A. Selective Colorimetric Detection of Polynucleotides Based on the Distance-Dependent Optical Properties of Gold Nanoparticles. *Science* **277,** 1078-1081 (1997).

[102] Le, J. D., Pinto, Y., Seeman, N. C., Musier-Forsyth, K., Taton, T. A. & Kiehl, R. A. DNA-templated self-assembly of metallic nanocomponent arrays on a surface. *Nano Lett.* **4,** 2343-2347 (2004).

[103] Morrison, L. E. & Stols, L. M. Sensitive Fluorescence-based thermodynamic and kinetic measurements of DNA hybridization in solution. *Biochemistry* **32,** 3095-3104 (1993).

[104] Panyutin, I. G. & Hsieh P. The kinetics of spontaneous DNA branch migration. *Proc. Nat. Acad. Sci.* **91,** 2021-2025 (1994).

[105] Simmel, F. C. & Yurke, B. A DNA-based molecular device switchable between three distinct mechanical states. *Applied Physics Letters* **80,** 883-885 (2002).

[106] Bishop, J. D. & Klavins, E. An improved autonomous DNA nanomotor. *Nano Lett.* **7,** 2574-2577 (2007).

[107] Ding, B. & Seeman, N. C. Operation of a DNA robot arm inserted into a 2D DNA crystalline substrate. *Science* **314,** 1583-1585 (2006).

[108] He, Y., Ye, T., Su, M., Zhang, C., Ribbe, A. E., Jiang, W. & Mao C. Hierarchical self-assembly of DNA into symmetric supramolecular polyhedra. *Nature* **452,** 198-201 (2008).

[109] Zheng, J., Birktoft, J. J., Chen, Y., Wang, T., Sha, R., Constantinou, P. E., Ginell, S. L., Mao, C. & Seeman, N. C. From molecular to macroscopic via the rational design of a self-assembled 3D DNA crystal. *Nature* **461,** 74-77 (2009).

[110] Yin, P., Hariadi, R. F., Sahu, S., Choi, H. M. T., Park, S. H., Labean, T. H. & Reif, J. H. Programming DNA Tube Circumferences. *Science* **321,** 824-826 (2008).

[111] Lincoln, T. A. & Joyce, G. F. Self-Sustained Replication of an RNA Enzyme. *Science* **323,** 1229-1232 (2009).

[112] Levy, M. & Ellington, A. D. Exponential growth by cross-catalytic deoxyribozymogens. *Proc. Nat. Acad. Sci.* **100,** 6416-6421 (2003).

Chapter 3: Kinetic Characterization of Strand Displacement Reactions

In every engineering discipline, there are some basic building blocks that are used over and over to great effect because they're just so darn reliable. Mechanical engineering has levers, pulleys, wheels, and ball bearings. Electrical engineering has resistors, capacitors, and inductors. In our field of DNA biomolecular engineering, particularly in creating dynamical systems, strand displacement is our building block of choice.

Strand displacement is the process of one single-stranded molecule of DNA (*strand*) displacing another from a multi-stranded complex. Although strand displacement had been studied in the 1970's and 1980's, Bernard Yurke's work on the DNA tweezers (*Nature,* 2000) was one of the first works to use strand displacement as a method for programming DNA interactions. In particular, Yurke showed that the kinetics of strand displacement can be greatly increased by a short single-stranded *toehold* next to the branch migration region. Since then, strand displacement based on toeholds have been used in a variety of ways to accomplish assorted nanoscale devices, such as logic gates and networks.

Although strand displacement had been established as a reliable tool, the kinetics of its function were not well characterized. The rate constant of strand displacement varied by 6 orders of magnitude depending on the toehold sequence and length. Researchers knew that stronger binding toehold usually corresponded to exponentially faster strand displacement

kinetics, but the quantitative relationship was not known.

The work below represents our effort to quantitatively model the strand displacement mechanism, and relate its kinetics to the binding energy of the toehold. To facilitate our modeling, we made sure to design all our sequences to have minimal secondary structure, so as to minimize kinetic folding and unfolding processes that may confound our observation of the strand displacement rates. In the end, our model was reduced to only 2 fitted parameters, and quantitatively predicted the rate constants of strand displacement to within a factor of 2 for 71 of the 85 different reactions we performed (the rest were all within a factor of 10). It is our hope and expectation that this work will serve as a user's manual for researchers intending to use strand displacement as a building block for designing more complex DNA devices and networks.

I came up with the initial idea for this work in Erik Winfree's laboratory in 2007, and collected kinetic data over the course of 2007. Erik and I then discussed the data and discussed various models of toehold exchange that would fit the observed data well. We argued and argued until we compromised on the three-step model presented here (I originally wanted a 2-step model, while Erik wanted a 4-step model). We then argued some more about how to properly calculate the binding energies of the toeholds. Much thanks goes to Xi Chen, Anne Condon, Niles Pierce, and Joseph Schaeffer, with whom we discussed our paper. Additionally, thanks go to Karthik Sarma, my 2009 SURF student, who caught a bug in my Matlab rate constant fitting code. This work was published in full as:

Zhang, David Yu; Winfree, Erik. "Control of DNA Strand Displacement Kinetics Using Toehold Exchange." *J. Am. Chem. Soc.* 131: 17303-17314 (2009).

DNA is increasingly being used as the engineering material of choice for the construction of nano-scale circuits, structures, and motors. Many of these enzyme-free constructions function by DNA strand displacement reactions. The kinetics of strand displacement can be modulated by toeholds, short single-stranded segments of DNA that co-localize reactant DNA molecules. Recently,

the toehold exchange process was introduced as a method for designing fast and reversible strand displacement reactions. Here, we characterize the kinetics of DNA toehold exchange, and model it as a three-step process. This model is simple and quantitatively predicts the kinetics of 85 different strand displacement reactions from the DNA sequences. Furthermore, we use toehold exchange to construct a simple catalytic reaction. This work improves the understanding of the kinetics of nucleic acid reactions, and will be useful in the rational design of dynamic DNA and RNA circuits and nanodevices.

DNA nanotechnology has emerged as a method of constructing structures [1–4], motors [5–9], and circuits [13–16] at the nanometer scale. Major reasons for the success of nucleic acids as nanoscale engineering materials are the predictability of their double-helical structure and of their Watson-Crick binding thermodynamics [17–32]. This predictability has allowed the rational design of remarkably complex static structures that self-assemble from synthetic DNA oligonucleotides of defined sequence [1–4]. Furthermore, DNA's programmable interactions can be used to mediate the molecular interactions of other materials, such as carbon nanotubes [33, 34] and gold nanoparticles [35, 36], and other chemistries [37–40].

Non-equilibrium DNA devices with dynamic function have also been constructed. These devices utilize DNA hybridization, branch migration, and dissociation for active state reconfiguration in which the kinetic pathways were designed [5, 8, 9, 16, 41]. Although the kinetics of DNA hybridization, branch migration, and dissociation processes have each been studied individually [32, 42–47], currently no method exists that accurately predicts the kinetics of DNA strand rearrangement processes from the nucleotide sequence. Establishment of a well-understood kinetic model of DNA interactions would further facilitate the widespread use of DNA devices as a technology for mediating nanoscale interactions with active behaviors.

Here, we specifically study one class of DNA reactions, that of strand displacement, in which one strand of DNA displaces another in binding to a third strand with partial complementarity to both. It is highly desirable, from an engineering perspective, to exactly

predict the kinetics of strand displacement reactions from only the strand and domain sequences, as this class of reactions has been used to construct a variety of DNA devices, including logic gates, [15, 48, 49], catalysts [16, 50–52], and motors [5, 8–12]. In these systems, a short single-stranded overhang region (known as a *toehold*) initiates the strand displacement reaction. Previous characterization of the kinetics of strand displacement reactions observed an exponential dependence of kinetics on the length of the toehold [53], but did not establish a general method to quantitatively predict kinetics from toehold sequence. The general solution to this problem is complicated by kinetics of hybridization nucleation and the unfolding of unintended secondary structures.

Our current work takes a step towards this goal by providing a three-step model of strand displacement that, under certain conditions, allows quantitative estimations of strand displacement rate constants from the thermodynamics of oligonucleotide hybridization [17, 20, 21] for secondary structure-free molecules. The rate constants predicted by our model were within one order of magnitude of the experimental best fit rate constants for all 85 different strand displacement reactions we ran for this paper, and within a factor of 2 for 71 of the 85 reactions that we tested. To further evaluate the predictive ability of the model for networks of reactions, we examine the kinetic behavior of a simple non-covalent DNA catalytic reaction network, again obtaining reasonable agreement with experiments. Remarkably, our model needed to fit values for only two parameters: rate constants for DNA hybridization and for branch migration. The former was fitted to three different values for three different classes of sequences.

The kinetics of nucleic acid strand displacement and hybridization are often confounded by the existence of unwanted secondary structure in strands or regions intended to be single-stranded [32, 54]. For the purposes of this paper, we wished to study the kinetics of strand displacement in isolation of such complications; thus, we carefully designed the experimental DNA sequences to possess minimal unwanted secondary structure. The model of strand displacement presented in this paper is expected to yield accurate rate constant predictions only when the underlying DNA sequence possesses relatively little unwanted secondary structure.

Domain Notation. In theory any guanine (G) can bind to any cytosine (C), but in practice the reverse reaction is so fast that the expected lifetime of an isolated one base-pair binding is very short. A stretch of several consecutive nucleotides must be complementary in order for the bound state to exist as an intermediate for further reaction at longer timescales. One useful abstraction for understanding hybridization-based constructions that exploit such intermediates is the *domain,* a consecutive stretch of nucleotides designed to act as a unit in binding. In this paper, domains are represented by Greek letters (Fig. 3-1A). Barred Greek letters denote domains complementary to the domains represented by unbarred Greek letters (e.g. $\overline{\beta}$ is complementary to β).

Because we perform experiments with a series of DNA molecules differing from each other only by a few bases, in this paper we further subdivide domains using superscript and subscript to denote the 5' and 3' portions of the domain, respectively. A superscript value of m indicates the 5'-most m bases of the full domain, while a subscript value of m indicates all but the 5'-most m bases of the domain. Thus the concatenation of β^m and β_m form the full β domain for all values of m.

Single-strand molecules of DNA (*strands*) consist of one or more concatenated domains, and are represented by capital English letters. DNA *complexes* are composed of several strands bound non-covalently to each other, and are also represented by capital English letters.

The domain abstraction simplifies the sequence design process for hybridization-based DNA devices. The interactions among DNA strands are determined by domain complementarities; the exact sequences of the domains have relatively little impact on the binding properties of the strands, except insofar as poor sequence design could lead to spurious binding between non-complementary domains. This sequence flexibility allows many instances of hybridization-based DNA systems to be constructed and run simultaneously.

Toeholds. When two strands share a common domain, they compete for binding to strands with complementary single-stranded domains. In Fig. 3-1B, strand X (green) and strand Y (red) both possess the β domain, and thus compete for binding to the $\overline{\beta}$ domain on the base strand (blue). Strand X, however, possesses an extra *toehold* domain γ, the

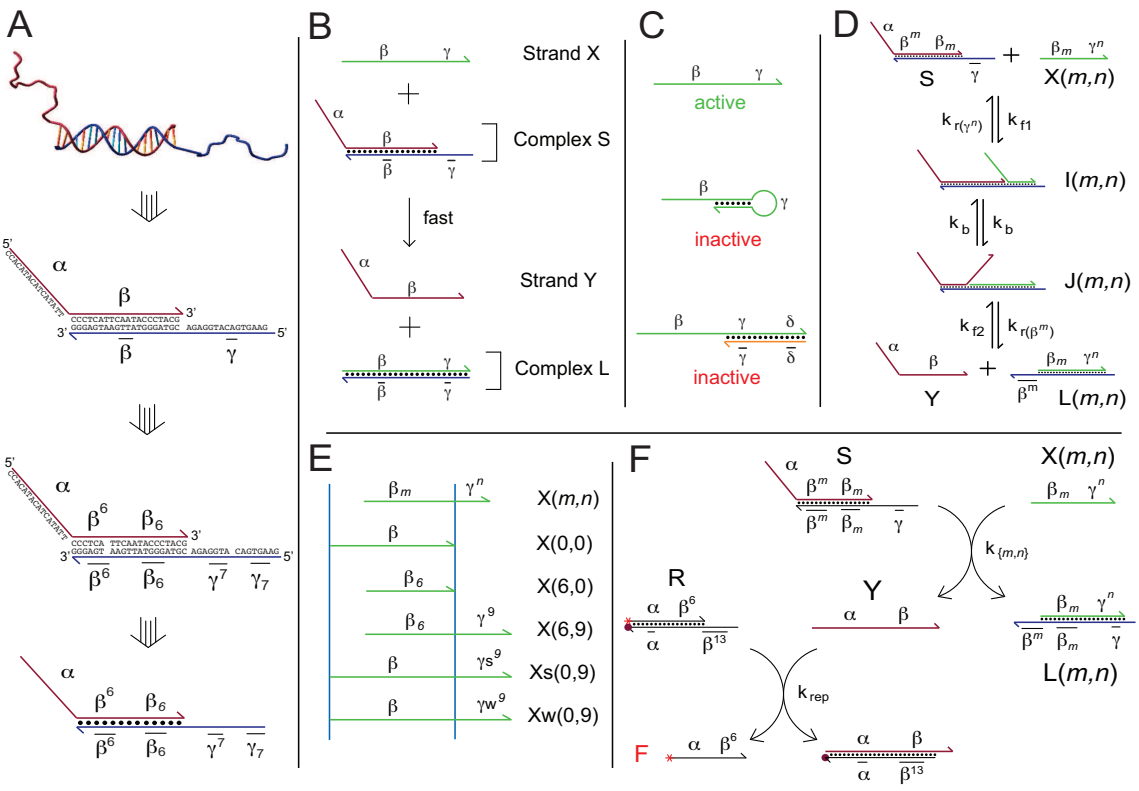


FIG. 3-1: **(A)** DNA abstraction. A DNA complex (top) is typically abstracted as several directional lines, one for each strand, with base identities shown. Here, we abstract DNA strands and complexes one step further by grouping contiguous nucleotides into domains, functional regions of DNA that act as units in binding. Because the principles and mechanisms studied in this paper are expected to be generalizable to most DNA sequences, we typically do not show the sequences of DNA strands in figures. For sequences, refer to Table 3-1. **(B)** A toehold mediated strand displacement reaction. The displacement of strand Y by strand X is facilitated by strand X's toehold domain γ. **(C)** Two examples of toehold sequestration. A strand of DNA can be rendered unreactive by inactivating its toehold domains. In the figure, toehold γ is sequestered through isolation in a hairpin (middle) and through hybridization to a complementary domain (bottom). **(D)** A toehold exchange reaction and its mechanism. Invading strand $X(m,n)$ binds to substrate complex S by toehold γ^n (known as the invading toehold) to form intermediate $I(m,n)$. Intermediate $I(m,n)$ represents all branch migration states in which Y is bound to more bases of $\overline{\beta_m}$ than $X(m,n)$. Intermediate $I(m,n)$ rearranges to form intermediate $J(m,n)$, which analogously represents all states in which $X(m,n)$ binds more bases of $\overline{\beta_m}$ than Y. Domain β^m (known as the incumbent toehold) spontaneously dissociates, releasing products Y and $L(m,n)$. The toehold exchange reaction is reversible, because strand Y can subsequently bind to complex $L(m,n)$ via strand Y's toehold β^m. **(E)** Comparison of various invading strands $X(m,n)$. Strand $X(m,n)$ is the concatenation of domains β_m and γ^n, and consequently has length $(b+n-m)$ nt, where b is the length of the full β domain. In a toehold exchange reaction using $X(m,n)$, the invading toehold has length n and the incumbent toehold has length m. For our experiments, we used three sets of invading toeholds, γ^n, γs^n, and γw^n. The sequence composition of the latter two are purely A/T's and purely G/C's, respectively, to characterize the kinetics of toehold exchange given weak and strong toeholds, respectively. Substrates using $\overline{\gamma s}$ are labeled Ss, and inputs using γs are labeled $Xs(m,n)$, and similarly for γ and γw. **(F)** Schematic of the experimental system used for rate measurements. Reporter complex R reacts stoichiometrically with product Y to yield increased fluorescence.

Dom.	Sequence	Length (nt)
α	5'- CCACATACATCATATT -3'	16
β	5'- CCCTCATTCAATACCCTACG -3'	$b \equiv 20$
γs	5'- CCCGCCGCCG -3'	10
γ	5'- TCTCCATGTCACTTC -3'	15
γw	5'- ATTTATTATA -3'	10

Table 3-1: Domain sequences

complement of which ($\overline{\gamma}$) also exists on the base strand. This toehold domain allows X to be co-localized to substrate complex S, even though strand Y is already bound. Branch migration of the β domain then allows strand X to displace strand Y. Strand Y possesses its own unique domain α, and can further react with other DNA complexes once released. This process is an example of *toehold mediated strand displacement,* in which a toehold facilitates a strand displacement reaction.

The reverse reaction, where strand Y displaces strand X from complex L, occurs at a rate up to six orders of magnitude slower, and therefore can be considered effectively unreactive. This is because the complex L does not possess a single-stranded complement to strand Y's α domain, so Y cannot be easily colocalized to complex L. Thus, the presence and properties of the toehold domain are instrumental to the kinetic control of DNA strand displacement reactions [5, 53].

Because of the toehold's role in initiating strand displacement reactions, strands can be rendered effectively unreactive if the toehold domain is made inaccessible by *toehold sequestering.* Toehold sequestering can be achieved in a number of ways, the two most common of which are hybridization of the toehold to a complementary domain [9, 15, 16] and isolation of the toehold in a short hairpin structure where helix formation is difficult [6, 52, 55] (see Fig. 3-1C). Programmed sequestering and subsequent exposure of toehold domains allows precise control of order and timing over the reactions, and has been used in conjunction with toehold mediated strand displacement to construct molecular motors [5, 8, 9], polymerization reactions [6, 9], catalytic reactions [9, 51, 52], and logic gates [15, 48, 49].

Recently, the *toehold exchange* mechanism was introduced as a method for improved control of strand displacement kinetics [16]. Toehold exchange is similar to toehold mediated strand displacement in that an invading strand (X) binds by a toehold to initiate branch

migration, but differs from the latter in that the incumbent strand (Y) possesses a unique toehold that must spontaneously dissociate for the reaction to complete. Expanding on the example strand displacement reaction in Fig. 3-1B, the toehold exchange reaction and mechanism that we study experimentally is illustrated in Fig. 3-1DE for an invading toehold of length n and an incumbent toehold of length m ($m, n > 0$): Strand $X(m, n)$ binds to complex S via invading toehold γ^n and displaces strand Y's β_m domain by branch migration. Strand Y's incumbent toehold β^m then spontaneously dissociates, yielding free strand Y and complex $L(m, n)$. The end result of the toehold exchange reaction is that the originally active toehold γ^n is sequestered while the formerly sequestered toehold β^m is activated. Thus, the active toehold is "exchanged" from γ^n to β^m. Note that the α domain and the overhang on the γ domain in Fig. 3-1D are not pertinent to the toehold exchange reaction; they are present in the figure to accurately reflect the experimental reaction (α is used to trigger a downstream fluorescent reporter, see Experimental System).

Functionally, toehold exchange offers two main advantages over toehold mediated strand displacement. First, the partially double-stranded product resulting from a toehold exchange reaction $L(m, n)$ possesses a single-stranded domain $\overline{\beta^m}$, allowing it to undergo further toehold mediated reactions. Thus in toehold exchange, two reactive reactants ($X(m, n)$ and S) yield two reactive products (Y and $L(m, n)$), rather than yielding just one as in toehold mediated strand displacement. This preservation of the number of active molecules, in combination with the reaction's reversibility, can be useful in larger reaction networks [56–58].

Second, toehold exchange weakens the coupling between the kinetics of strand displacement and the thermodynamics of the reaction. In toehold mediated strand displacement reactions operating below their maximum rate, there is a strong coupling between the kinetics and the thermodynamics: to speed up the strand displacement reaction, the invading toehold must be made stronger thermodynamically, for example by increasing the toehold length [53]. Consequently, faster strand displacement reactions are more thermodynamically favorable in the net reaction. (The reaction kinetics of toehold mediated strand displacement saturates for sufficiently long toeholds, breaking the coupling between ther-

modynamics and kinetics.) In contrast, in toehold exchange, when both the invading and the incumbent toeholds (γ^n and β^m) are made stronger, the kinetics of both the forward and the reverse reactions are sped up, within the limits of sufficiently low concentration. Thus, a strand displacement reaction based on toehold exchange can be fast despite being only weakly thermodynamically favorable, or even thermodynamically unfavorable.

Toehold mediated strand displacement and toehold exchange encompass much larger classes of reactions than the examples presented in Fig. 3-1B and Fig. 3-1D. Both the substrate S and the input $X(m,n)$ may be part of larger DNA complexes (as in ref. [16, 56, 58]) or may be functionalized to other materials. As one example, inputs may be expressed on the surface of gold nanoparticles [35], while substrates are regularly arrayed on a DNA origami [2, 59], to create an autonomous moving device analogous to those constructed from deoxyribozymes [13]. Here, we experimentally characterize only reactions of the type presented in Fig. 3-1D, but it is expected that the results on the kinetics of toehold exchange can be generalized to a wide range of molecular designs.

Three-step model of toehold exchange. We simplify the biophysics of the toehold exchange process to the three-step model shown in Fig. 3-1D:

$$X(m,n) + S \underset{k_{r(\gamma^n)}}{\overset{k_{f1}}{\rightleftharpoons}} I(m,n) \underset{k_b}{\overset{k_b}{\rightleftharpoons}} J(m,n) \underset{k_{f2}}{\overset{k_{r(\beta^m)}}{\rightleftharpoons}} Y + L(m,n)$$

The rate constants k_{f1} and k_{f2} denote the hybridization rates of γ^n and β^m to their complements, respectively. For simplicity, here we assume that these two rate constants are equal in value, $k_{f1} = k_{f2} = k_f$. Later, we see that the base compositions of the γ^n and β^m domains can cause the values of k_{f1} and k_{f2} to vary significantly [71].

The value of $k_{r(\gamma^n)}$ denotes the first-order rate at which the toehold γ^n dissociates, and is calculated so as to satisfy the equilibrium between $X(m,n) + S$ and $I(m,n)$:

$$k_{r(\gamma^n)} = k_{f1} \cdot \frac{2}{b-m} \cdot e^{\Delta G^\circ(\gamma^n)/RT} \tag{1}$$

where $\Delta G^\circ(\gamma^n) < 0$ is the binding energy between γ^n and its complement, b is the length of the full β domain, and $b-m$ is the length of the β_m domain (since the state $I(m,n)$ contains

$\frac{b-m}{2}$ iso-energetic branch migration intermediate states). For the systems we examine, $b-m$ varies from 13 to 20 nt. The $\frac{2}{b-m}$ term is necessary to ensure the correct relative concentrations of the states at equilibrium [72]. The value of $k_{r(\beta^m)}$ is calculated similarly. Finally, k_b represents the effective rate at which the branch migration junction crosses the middle of β_m, and depends on the length of the branch migration domain β_m.

To quantitatively test the rate constants predicted by our model, we collected kinetic data on a series of 85 different strand displacement reactions. To generate the 85 different predicted rate constants, our model required values for k_{f1}, k_{f2}, and k_b, in addition to the generally accepted nucleic acid hybridization thermodynamics parameters. Two of these parameters, k_{f1} and k_b, were fitted to our experimental data, while the last (k_{f2}) was assumed to be the same as k_{f1}. Note that three different values of k_{f1} were fitted, corresponding to the three different γ domains experimentally tested. We expect these parameter values to be roughly the same for systems with similar salt concentrations, temperatures, and domain lengths and compositions; consequently we expect our model to be able to roughly predict the kinetics of DNA hybridization-based systems and circuits that utilize similar strands and complexes. Our predictions may not be accurate for strands that possess significant unwanted secondary structure.

Bimolecular reaction model of toehold exchange. For ease of designing, modeling, simulating, and data fitting, it is convenient to model toehold exchange and toehold mediated strand displacement as simple bimolecular reactions. In fact, the kinetics of toehold exchange and toehold mediated strand displacement have been previously observed to be well-approximated by bimolecular reactions with second-order rate constants [16, 53]. In terms of the example system shown in Fig. 3-1D,

$$X(m,n) + S \underset{k_{(\gamma^n,\beta_m,\beta^m)}}{\overset{k_{(\beta^m,\beta_m,\gamma^n)}}{\rightleftharpoons}} Y + L(m,n)$$

The associated forward and reverse second-order rate constants $k_{(\beta^m,\beta_m,\gamma^n)}$ and $k_{(\gamma^n,\beta_m,\beta^m)}$ for this bimolecular reaction model (BM) of strand displacement are dependent on the sequences of the β^m, β_m, and γ^n domains, as well as the ambient temperature and salt

concentrations.

Using our three-step model of toehold exchange, we can analytically derive the BM rate constant under quasi-steady state (QSS) conditions (see Text 3-S1):

$$k_{(\beta^m,\beta_m,\gamma^n)} \equiv \frac{k_{r(\beta^m)}k_f k_b}{k_{r(\gamma^n)}k_{r(\beta^m)} + k_{r(\gamma^n)}k_b + k_{r(\beta^m)}k_b} \tag{2}$$

The above equation yields the expression for the BM rate constant, but there is a subtle inaccuracy to it. The bimolecular reaction model concentrations $[X(m,n)]_{BM}$ and $[S]_{BM}$ differ from their three-step model counterparts $[X(m,n)]$ and $[S]$ because the BM does not account for intermediate concentrations $[I(m,n)]$ and $[J(m,n)]$. In order for the BM model and rate constant to accurately describe the kinetics of strand displacement processes, the concentrations of $X(m,n)$ and S must be sufficiently low. For example, the kinetics of Y production are accurate to within 50% when the initial concentrations of $[X(m,n)]$ and $[S]$ are below

$$c_{crit} = \frac{0.1}{k_f} \cdot \frac{k_{r(\gamma^n)}k_{r(\beta^m)} + k_{r(\gamma^n)}k_b + k_{r(\beta^m)}k_b}{k_b + k_{r(\beta^m)}} \tag{3}$$

The 0.1 in the numerator of the first term varies non-linearly with the required accuracy of the BM model (with larger numbers leading to higher inaccuracy). The derivation of this critical concentration is shown in Text 3-S2. Within these concentration limits, the BM rate constants for a particular system can be derived from the parameters of the three-step model, which can in turn be predicted based on the sequences of the strands.

The experimental system. A typical fluorescence kinetics experiment contains substrate complex S, input strand $X(m,n)$, and fluorescent reporter complex R mixed in solution (Fig. 3-1F), with the input $X(m,n)$ added last to trigger the start of the reaction. $X(m,n)$ reacts with S to form Y and $L(m,n)$, the former of which undergoes further reaction with reporter R to yield an increased fluorescence signal. In addition to allowing high-resolution kinetic data, the reporter complex R reduces the back-reaction of Y with

product complex $L(m,n)$ by quickly removing free Y from solution. Another reason for using a separate reporter complex rather than directly labelling X or S with fluorophores and quenchers is the extent to which fluorophore-quencher binding alters the thermodynamics [66]. The indirect reporter complex R isolates this effect from toehold exchange thermodynamics and kinetics. When bound in S, strand Y does not react with R because the relevant toehold is sequestered. Strand $X(m,n)$ does not react significantly with R because it lacks the α domain.

Three different versions of S and $X(m,n)$ were used, corresponding to three different sequences for domain γ. The default sequence, which we refer to as γ with no further annotations, is representative of toeholds with typical binding energies for their lengths; it consists of a mixture of G/C/A/T bases with all subdomains γ^n containing roughly equal numbers of G/C and A/T base pairs (e.g. γ^8 contains 4 A/T bases and 4 G/C bases). For weak toeholds and strong toeholds, we used domains γw and γs, which consist exclusively of A/T and G/C base pairs, respectively (see Table 3-2). For these three sequence choices, the corresponding strands and complexes are referred to as $X(m,n)$, S, and $L(m,n)$ for the typical toeholds, $Xw(m,n)$, Sw, and $Lw(m,n)$ for the weak toeholds, and $Xs(m,n)$, Ss, and $Ls(m,n)$ for the strong toeholds.

The domain sequences were carefully designed to avoid secondary structure in single-stranded species so that the kinetics of toehold exchange could be separated from the first-order kinetics of unfolding unintended secondary structures. To this end, one useful sequence design heuristic was to minimize the number of G's in single-stranded species (e.g. $X(m,n)$ and Y). Not only are unintentional G-C bases much more stable than A-T ones, G's also further contribute to undesired secondary structures through G-T wobbles and G-quartets [32]. For this reason, there are no G's in the α domain and 1 G each in the β and γ domains.

To systematically characterize the kinetics of toehold mediated strand displacement and toehold exchange, we measured the BM rate constants of the reaction between substrates Ss, S, and Sw and a series of different inputs $Xs(m,n)$, $X(m,n)$, and $Xw(m,n)$ for various values of m and n. Double-stranded waste products $Ls(m,n)$, $L(m,n)$, and $Lw(m,n)$

possess 5' overhangs on the bottom strand whenever n is smaller than the length of the corresponding γ domain. For convenience, the BM rate constants for forward reactions ($k_{(\beta^m,\beta_m,\gamma^n)}$ by the previous convention) are relabelled $ks_{\{m,n\}}$, $k_{\{m,n\}}$, and $kw_{\{m,n\}}$ for the rate constants using the three sets of sequences presented in this paper. For discussions not specific to sequence, we use the notation for typical toehold strengths, e.g. $k_{\{m,n\}}$.

Materials and Methods

DNA sequences and design. The sequences presented here are based on those of Zhang et al. [16], and were designed by hand to possess minimal secondary structure and crosstalk (binding between unrelated domains). NUPACK [17] calculates there to be no more than 4 paired bases between any pair of domains at 25 °C, even at 1 μM concentration. Thus, the domains we use can be approximated as structure-free. Similar sequences could have been designed using computer-aided methods [63]. Substantial secondary structure is known to slow down branch migration and interfere with hybridization [54].

Toehold binding energy calculations. The value of the predicted BM rate constant depends sensitively on the toehold binding energies $\Delta G^\circ(\gamma^n)$ and $\Delta G^\circ(\beta^m)$. We calculated each these as difference in standard free energies of two relevant complexes:

$$\begin{aligned}\Delta G^\circ(\gamma^n) &= \Delta G^\circ(I(0,n)) - \Delta G^\circ(S)\\ \Delta G^\circ(\beta^m) &= \Delta G^\circ(J(m,0)) - \Delta G^\circ(L(m,0))\end{aligned}$$

The calculation of the standard free energies of each complex used the initiation free energy and base stacking energy parameters reported by Santa Lucia et al. [21], and dangling end parameters reported by Bommarito et al. [22]. Additionally, we needed three energy parameters that have not been conclusively established in literature: the correction for DNA in a solution containing magnesium ions, the free energy contribution of the coaxial stack at a nick, and the contribution of "coaxial stack dangles", dangles adjacent to two coaxial helices (e.g. the β domain in state $I(m,n)$ in Fig. 3-1D).

Several different models for the energetics of DNA hybridization in magnesium solutions have been proposed and investigated [23–25], but it was unclear which model would best apply to our DNA complexes. However, all of these models suggested that the energetics of DNA hybridization were similar in a 1 M Na^+ solution as in a 10 mM Mg^{2+}. Consequently, for the calculation of the default toehold binding energies presented in the main paper, we simply used the energetics parameters for DNA hybridization in a 1 M Na^+ solution. Application of Owczarzy et al.'s magnesium correction formula [25] yielded standard free energies that were slightly less energetic than reported in the main paper (see Text 3-S3). The BM rate constants predicted using these modified parameters showed similar quality of fit (see Fig. 3-S2D).

Similarly, several different sets of coaxial stacking energetics parameters have been reported [23, 26–28]. In some cases, these values differed from each other by up to 2.5 kcal/mol. Using the values reported by Protozanova et al. [26], with minor adjustments for temperature [30, 31] (see Text 3-S3), yielded the best agreement between model-predicted BM rate constants and experimental best-fit rate constants. Using values reported by Pyshnyi et al. [27] yielded similar energetics for S and Ss, but significantly stronger binding energetics for Sw (by about 0.8 kcal/mol), and lead to slightly worse quality of fit for BM rate constants (see Fig. 3-S2A). Furthermore, Pyshnyi et al. [29] later reported that the energetics of coaxial stacking near a nick depends significantly on the nearest neighbors (bases one away from the nick). Note that several publicly available and commonly used DNA folding programs, such as mFold [19] and NUPACK [17], by default do not include any coaxial stacking term for their energetics calculations.

To the best of our knowledge, the energetics of coaxial stack dangles has not been comprehensively characterized. DNA folding programs such as mFold and NUPACK by default use the energetics parameters of "terminal dangles" (those adjacent to a single helix, reported by Bommarito et al. [22]) as a substitute, but there is no *a priori* reason to believe that the energetics of coaxial stack dangles would be correlated with the energetics of terminal dangles. For our default energetics calculations, we assumed that the ΔG° contribution from these coaxial stack dangles is 0. Using the energetics parameters of

Toehold	Binding Energy	Toehold	Binding Energy
γ^0	+1.9	γs^0	+1.9
γ^1	+0.2	γs^1	-1.1
γ^2	-1.7	γs^2	-3.2
γ^3	-3.0	γs^3	-5.0
γ^4	-4.7	γs^4	-8.0
γ^5	-6.9	γs^5	-10.3
γ^6	-8.3	γs^6	-12.1
γ^7	-9.2	γs^7	-15.1
γ^8	-11.9	γs^8	-17.3
γ^9	-12.9	γs^9	-19.2
γ^{10}	-14.8	γs^{10}	-21.2
γ^{15}	-21.8		
β^0	+1.2	γw^0	+1.9
β^1	-0.6	γw^1	+0.2
β^2	-2.7	γw^2	-0.8
β^3	-4.5	γw^3	-2.1
β^4	-5.6	γw^4	-3.8
β^5	-6.7	γw^5	-4.3
β^6	-9.5	γw^6	-5.3
β^7	-10.2	γw^7	-7.0
		γw^8	-7.5
		γw^9	-9.0
		γw^{10}	-8.9

Table 3-2: Calculated toehold binding energies (in kcal/mol)

terminal dangles instead of 0 for coaxial stack dangles led to slightly worse quality of fit for BM rate constants (see Fig. 3-S2BC).

The default toehold binding energies (as described above using works by Santa Lucia et al. [21], Bommarito et al. [22], and Protozanova et al. [26]) are reported in Table 3-2. These values were the ones used to generate the model-predicted BM rate constants in Figs. 3-3, 3-4, and 3-6. A detailed step-by-step procedure of how to derive the toehold binding energies is shown in Text 3-S3.

The uncertainty of the absolute values of the toehold binding energies may be 1 kcal/mol or more, but uncertainty in the relative binding energies of toeholds of different lengths is smaller, because the latter depend primarily on the well-established energy parameters for stacks and terminal dangles. A consequence of the former, however, is that the value of parameter k_b is known to only an order of magnitude. Using some of the alternative energetics parameters discussed above, the fitted value of k_b range from 0.5 s^{-1} to 7 s^{-1}.

Buffer conditions. DNA oligonucleotides were stored in TE buffer (10 mM Tris · HCl pH balanced to 8.0, with 1 mM EDTA·Na_2, purchased as 100x stock from Sigma-Aldrich) at

4 °C. Directly preceding experiments, TE buffer with 62.5 mM $MgCl_2$ was added at 1:4 ratio to the sample, achieving a final $MgCl_2$ concentration of 12.5 mM (of which 1 mM is bound to EDTA). This buffer is henceforth known as "TE/Mg^{2+}" buffer. All experiments and purifications were performed at 25±0.5 °C, with temperature controlled using an external temperature bath.

Substrate purification. DNA oligonucleotides used in this study were purchased from Integrated DNA Technologies (IDT), with HPLC purification. Where applicable, fluorophores were attached by IDT as well. Concentrations were determined from the measured absorbance at 260 nM using an Eppendorf Biophotometer and the calculated extinction coefficients [65].

Substrate complexes S, Ss, and Sw, and the reporter complex R were further purified by non-denaturing (ND) polyacrylamide gel electrophoresis (PAGE) as follows: Strands for each sample were prepared with nominally correct stoichiometry at 20 μM and annealed. The samples were then run on 12% ND PAGE at 180 V for 6 hours.

The acrylamide (19:1 acrylamide:bis) was diluted from 40% acrylamide stock (Ambion). ND loading dye containing xylene cyanol FF in 50% glycerol was added to all samples, achieving final gycerol concentration of 10% by volume. Gels were run at 25 °C using a Novex chamber with external temperature bath.

The proper bands were cut out and eluted in 2 mL TE/Mg^{2+} buffer for 2 days. Purified complexes were quantitated by measurement of absorbance at 260 nm using an Eppendorf Biophotometer, and calculated extinction coefficients [65]. Typical yields ranged from 40% to 60%.

Annealing. All annealing processes were performed with an Eppendorf Mastercycler Gradient thermocycler. The samples were brought down from 95 °C to 20 °C at a constant rate over the course of 90 minutes.

Spectrofluorimetry studies. Spectrofluorimetry studies were done using a SPEX Fluorolog-3 (Horiba) with 1.6 mL 119-004F synthetic quartz cells (Hellma). The excitation was at 588 nm, while emissions was at 602 nm (optimal signal for ROX fluorophore). In all spectrofluorimetry experiments, the total reaction volume was 1.5 mL. For net reaction

studies in which the concentration of the reporter was in excess of 1 nM (see Tables 3-S2, 3-S3, and 3-S4), 2 nm band pass slits were used for both excitation and emission monochromators; for experiments where reporter concentration was less than 1 nM, 4 nm slits were used. Experiments shown in Figs. 3-2D, 3-3A, 3-4A, 3-5B, and 3-7B, as well as experiments not shown, were done with integration time of 10 seconds for every 60 second time-point. The experiment shown in Fig. 3-7A was done with integration time of 10 seconds for every 10 second time-point.

Prior to each experiment, all cuvettes were cleaned thoroughly: each cuvette was washed 15 times in distilled water, once in 70% ethanol, another 5 times in distilled water, and finally once more in 70% ethanol.

For the slit size, concentrations, and times chosen, no measurable photobleaching was observed. Fluorescence measurements are linear in the concentration of the free fluorescent strand F. All experimental results were within the linear regime of the spectrofluorimeter detector, according to specification sheets provided by the manufacturer.

Fluorescence normalization. Fluorescence is normalized so that 1 normalized unit (n.u.) of fluorescence corresponds to 1 nM of unquenched fluorophore-labelled strand F. This normalization is based on the fluorescence levels of annealed samples: a negative control with $[R] = 30$ nM, $[X(0,10)] = 20$ nM and $[S] = 0$ nM, and a positive control with $[R] = 30$ nM, $[X(0,10)] = 20$ nM and $[S] = 10$ nM.

Day-to-day and sample-to-sample variations are estimated to be less than 5% [16].

Carrier strands. In our procedures, DNA sticks non-specifically to pipette tips, so that serial dilutions lead to stocks more dilute than expected [16]. Unfortunately, this loss is not consistent, so we could not compensate for tip loss with additional reagent. Instead, we introduced into all dilute stocks (1 μM and below) a non-reactive 20 nt poly-T "carrier" strand, at a concentration of 1 μM. Because pipette tip loss is non-specific, the majority of DNA loss would be of the carrier strand, and serially diluted stocks are only slightly more dilute than expected. Poly-T strands have minimal influence on the reactions of other DNA molecules in this system [16].

Parameter Fitting. The best-fit rate constants to experimental data were fitted using

the 'fminunc' function in Matlab to minimize the error between experimental data and the reaction model. Sample code is shown in text 3-S4. The error is calculated as follows:

$$\text{Error} = \sum_{t,\text{traces}} \frac{(F_d(t) - F_m(t))^2}{F_m(t)}$$

where $F_d(t)$ is the fluorescence value of the data at time t, and $F_m(t)$ is the fluorescence value predicted by the ODE model at time t. The sum of square error is divided by the expected fluorescence value $F_m(t)$ because the fluorescence readings yielded by the photomultiplier tube follow a Poisson distribution.

Rate constants were fitted one at a time; in every case, all but one rate constant were previously fitted. For example, the rate constant $k_{rep} = 1.3 \cdot 10^6$ M^{-1} s^{-1}, established by experimentation and fitting (Fig. 3-2), was used to fit the value of $k_{\{0,5\}}$. Matlab yielded $k_{\{0,5\}} = 1.0 \cdot 10^6$ M^{-1} s^{-1} as the best-fit value of $k_{\{0,5\}}$ for the following reaction network to fit the data in Fig. 3-3A:

$$S + X(0,5) \overset{k_{\{0,5\}}}{\rightarrow} Y + L(0,5)$$
$$Y + R \overset{k_{rep}}{\rightarrow} F$$

Errors in the quantitation of DNA oligonucleotide concentrations can be reflected as errors in the inferred BM rate constants. For example, DNA strands and complexes being quantitated to be 5% more concentrated than their actual concentration would lead to an inferred rate constant 5% lower than the actual value.

Error bars shown in Fig. 3-3BC were generated using "leave one out" error: best-fit rate constants were generated for each pair of data traces (e.g. the 0.2 nM $X(0,5)$ and the 0.4 nM $X(0,5)$ traces of Fig. 3-3A), and the standard deviation of these three pairs is calculated. The error bars show two standard deviations above and below the rate constant fitted using all three data traces. For subsequent figures, no error bars are plotted because each data point represents the best-fit BM rate constant to a single experimental trace.

For short invading toehold lengths ($n < 3$), the initial data points with fluorescence less

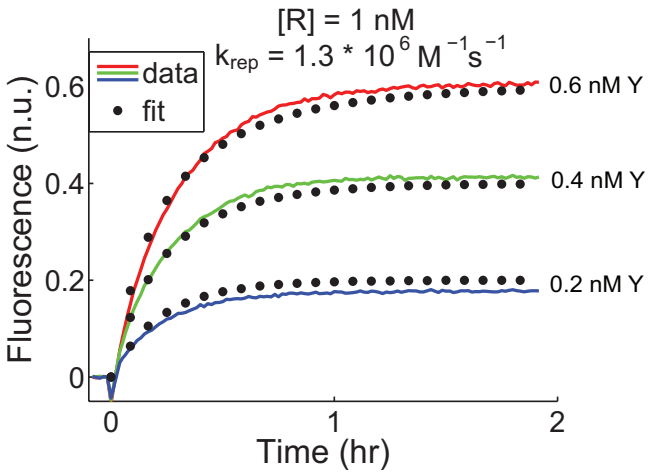


FIG. 3-2: Kinetic characterization of the reporter complex R. "n.u." denotes normalized units; all fluorescent results were normalized so that 1 n.u. corresponds to 1 nM of free F. Reporter R was initially in solution at the indicated concentration, and Y was added to solution at $t \approx 0$ to achieve the final concentration displayed. Dotted line shows simulation traces of a second-order displacement reaction with rate constant $k_{rep} = 1.3 \cdot 10^6$ /M /s.

than 1 nM were ignored in evaluating the error score. This is because in these experiments, there was an initial fast reaction that very quickly yielded about 0.5 nM of F; this fast reaction is assumed to be due to synthesis errors in the oligonucleotides, and thus is not representative of the reaction of interest.

Results and Discussion

Reporter complex characterization.

Fig. 3-2 shows the kinetics of the reaction between product Y and reporter complex R. A second order reaction with rate constant $k_{rep} = 1.3 \cdot 10^6$ M^{-1} s^{-1} fits the data well at the concentrations tested; this parameter value was used for all subsequent modeling.

Toehold mediated strand displacement, $m = 0$.

The rate constant $k_{\{0,5\}} = 1.0 \cdot 10^6$ M^{-1} s^{-1} yielded the best fit to the experimental data in Fig. 3-3A. Similar experiments and simulations were done for other inputs $X(0,n)$, $Xs(0,n)$, and $Xw(0,n)$ (data not shown), and all best-fit rate constants $k_{\{0,n\}}$ showed similar or better quality of fit to experimental data. The concentrations of reporter, substrate, and input used for each set of experiments are listed in Tables 3-S2, 3-S3, and 3-S4. The concentrations were chosen so that the experiments finished within a reasonable amount of time (between

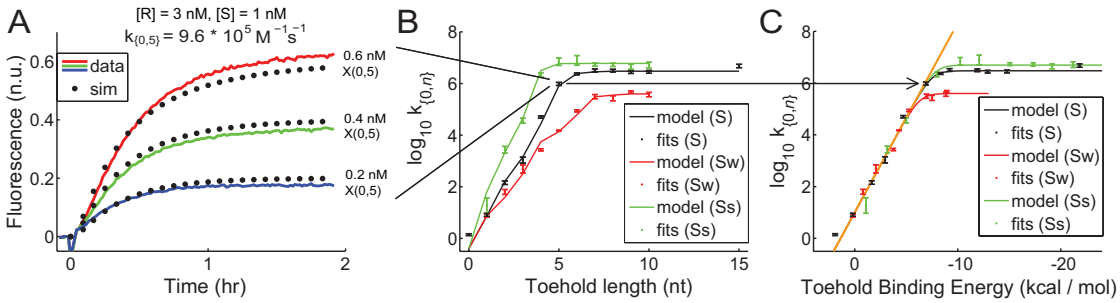


FIG. 3-3: Toehold mediated strand displacement experiments ($m = 0$). Results presented in this figure also use domains γs and γw in place of domain γ where specified. **(A)** Sample trajectories for $n = 5$. S and R were initially in solution at the displayed concentrations, and $X(0, 5)$ was added to solution at $t \approx 0$ to achieve the final concentration displayed. The black dotted lines labelled "fit" denote simulations of a bimolecular reaction with with the experimental best-fit rate constant $k_{\{0,5\}} = 1.0 \cdot 10^6$ M^{-1} s^{-1}, and the reporter reaction with rate constant $k_{rep} = 1.3 \cdot 10^6$ M^{-1} s^{-1}. The black lines spanning (A), (B), and (C) indicate that the three traces shown in (A) are represented by a single data point in (B) and (C). **(B)** Summary of strand displacement rate constants plotted against the invading toehold length n. The y-axis shows the base-10 logarithm of the experimental best-fit values ("fits") and three-step model predicted values ("model") of $ks_{\{0,n\}}$, $k_{\{0,n\}}$, and $kw_{\{0,n\}}$, expressed in M^{-1} s^{-1}. **(C)** Summary of strand displacement rate constants plotted against the calculated binding energy of the toehold. The orange line shows the asymptotic behavior predicted by the three-step model as k_f approaches ∞.

10 minutes and 3 days). Subsequent analysis showed that the concentrations were below the critical concentration c_{crit} derived earlier.

The best-fit values of $ks_{\{0,n\}}$, $k_{\{0,n\}}$, and $kw_{\{0,n\}}$ are plotted semi-logarithmically against the length of the toehold n in Fig. 3-3B (small dots with error bars), and against the predicted toehold binding energies in Fig. 3-3C. In both, the experimental best-fit rate constants follow a "kinked line" distribution, with an initial sloped region and an eventual flat region.

This kinked line behavior is predicted by our three-step model (shown as solid lines in Fig. 3-3BC). Recall equation (2), the expression for the BM rate constant of toehold exchange:

$$\begin{aligned} k_{\{m,n\}} &\equiv k_{(\beta^m,\beta_m,\gamma^n)} \\ &= \frac{k_f k_{r(\beta^m)} k_b}{k_{r(\gamma^n)} k_{r(\beta^m)} + k_{r(\gamma^n)} k_b + k_{r(\beta^m)} k_b} \end{aligned}$$

For the toehold mediated strand displacement experiments in this section, $m = 0$, so $k_{r(\beta^0)} >> k_b$, and the $k_{r(\gamma^n)} k_b$ term can be eliminated from the denominator because it

is dominated by $k_{r(\gamma^n)}k_{r(\beta^0)}$ term.

In the flat region, the binding energy of the invading toehold is strong enough that $k_b > k_{r(\gamma^n)}$, and the expression can be approximated as:

$$k_{\{0,n\}} \equiv k_{(\beta^0,\beta_0,\gamma^n)} \approx \frac{k_f k_{r(\beta^0)} k_b}{k_{r(\beta^0)} k_b} = k_f \tag{4}$$

Different values of k_f were fitted to the three different sets of traces (with S, Ss, or Sw). In general, higher G-C content appears to increase the hybridization rate of complementary strands (k_f). For the substrate and inputs using the γs^n toehold, the best-fit value of k_f was the highest at $6 \cdot 10^6$ M^{-1} s^{-1}. The values $k_f = 3 \cdot 10^6$ M^{-1} s^{-1} and $k_f = 4 \cdot 10^5$ M^{-1} s^{-1} fit best for substrates and inputs with the γ and γw toeholds, respectively.

In the sloped region, $k_{r(\gamma^n)} > k_b$, and we can approximate the BM rate constant expression as:

$$k_{\{0,n\}} \approx \frac{k_f k_{r(\beta^0)} k_b}{k_{r(\gamma^n)} k_{r(\beta^0)}} = \frac{k_f k_b}{k_{r(\gamma^n)}} \tag{5}$$

$$= k_b \frac{b-m}{2} e^{-\Delta G^\circ(\gamma^n)/RT} [\mathrm{M}^{-1}] \tag{6}$$

Taking the logarithm (base 10) of the BM rate constant,

$$\log_{10}(k_{\{0,n\}}) \approx \log_{10}(\frac{k_b(b-m)}{2}) - \frac{\log_{10} e}{RT} \Delta G^\circ(\gamma^n) \tag{7}$$

It can be seen that the slope and y-intercept of Fig. 3-3B are predicted by the three-step model to be $\frac{\log_{10} e}{RT} \approx 0.732$ and $\log_{10}(\frac{k_b(b-m)}{2})$, respectively. Taking the 0.732 slope as a given, the best-fit value of the y-intercept is calculated by Matlab to be 1.0. For toehold mediated strand displacement, the length of the branch migration is $b-m = 20-0 = 20$ nt, so the corresponding best-fit value of k_b is 1.0 s^{-1}. This value of k_b fit well for all three sets of experiments as expected, because k_b should be dependent only on the branch migration domain β_m, which is the same for all three sets of experiments.

We do not currently have a good method of estimating k_f from first principles: The γs toehold possesses a k_f about twice as large as the γ toehold, and about twenty times as large as the γw toehold. Thus, strong base-pairs seem to have a disproportionate effect on determining k_f. It is likely that k_f reflects the biophysics of hybridization nucleation [32,

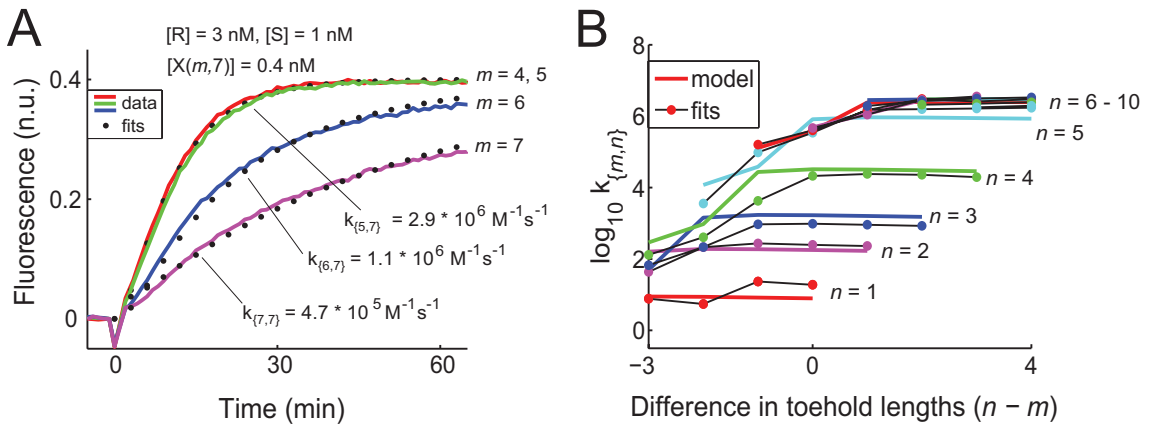


FIG. 3-4: Toehold exchange. All results presented in this figure use only domain γ (and not γs or γw). **(A)** Sample trajectories for $n = 7$ for various m. R and S were present in solution initially, and $X(m, 7)$ was added at $t \approx 0$. The black dotted lines labelled "fit" denote simulations of a bimolecular reaction with the experimental best-fit rate constants $k_{\{m,n\}}$ shown in the figure. **(B)** Summary of toehold exchange rates. The solid lines labelled "model" show the BM rate constants $k_{\{m,n\}}$, with $k_f = 3 \cdot 10^6$ M^{-1} s^{-1}, and $k_b = 1.0 \cdot (\frac{b}{b-m})^2$ s^{-1}. These predictions were not fitted to the data in this figure; they depended only on the values of k_f and k_b fitted from the earlier experiments on toehold mediated strand displacement. The dots labelled "fits" show the experimental best-fit rate constants. Unlike in Fig. 3-3, only one concentration of each $X(m, n)$ was tested here.

60, 69].

Toehold exchange, $m > 0$.

Experiments to characterize the kinetics of toehold exchange are similar to those of toehold mediated strand displacement. Fig. 3-4A shows the data and simulations using the best-fit $k_{\{m,n\}}$ [73] for the experiments using the set of inputs $X(m, 7)$, where m ranges between 4 and 7. The best-fit $k_{\{m,n\}}$ are plotted as dots in Fig. 3-4B, and compared against the values of $k_{\{m,n\}}$ predicted by our three-step model, using the value $k_f = 3 \cdot 10^6$ M^{-1} s^{-1}. The value of k_b is calculated as $1.0 \cdot (\frac{b}{b-m})^2$ s^{-1} to account for changes in the length of the branch migration region, in accordance with previous unbiased random walk models of branch migration kinetics [42, 43]. Notably, no new model parameters were fitted to generate the model-predicted rates in Fig. 3-4B.

The toehold exchange experimental results showed that the predicted BM rate constants agreed reasonably well to the experimental best-fit rate constants (with most values correct to within a factor of 2, and all values correct to within a factor of 10). Observed differences between the model-predicted and best-fit rate constants are at least partially due to errors

in the binding energies of the toeholds. Unintended secondary structures, particularly in the cases of Xs, Ss, Xw, and Sw, may have complicated the kinetics of strand displacement. Finally, strand synthesis impurities and inaccuracies in DNA extinction coefficients may have caused error in DNA quantitation, that would have been reflected as an error in observed rate constants.

The BM rate constants of toehold exchange ($k_{\{m,n\}}$) were similar in value to those of toehold mediated strand displacement $k_{\{0,n\}}$ when $m \leq n$. When $m > n$, the rate constant of toehold exchange decreases sharply with increasing m, both experimentally and as predicted by the model. This can be understood intuitively: Below the critical concentration, the initial association of $X(m,n)$ and S by the toehold is the rate limiting step. When an input molecule $X(m,n)$ binds to the substrate S, the resulting complex can dissociate either into $Y + L$, or into $X(m,n) + S$. The probability of the former case is determined by the relative binding energies of the incumbent and invading toeholds, and is evaluated (using equation (1)) as:

$$\Pr(I \rightarrow Y + L) \approx \frac{k_{r(\beta^m)}}{k_{r(\beta^m)} + k_{r(\gamma^n)}} \tag{8}$$

$$= \frac{e^{\Delta G^\circ(\beta^m)}}{e^{\Delta G^\circ(\beta^m)/RT} + e^{\Delta G^\circ(\gamma^n)/RT}} \tag{9}$$

As $n - m$ increases, the probability of dissociating into $Y + L$ approaches 1, and the BM rate constant approaches the hybridization rate of the toehold k_f. As $n - m$ decreases, the probability of dissociating into $Y+L$ and the BM rate constant decays roughly exponentially.

Catalysis based on toehold exchange.

From the previous section, the kinetics of toehold exchange are seen to approach that of toehold mediated strand displacement (m=0) when $n \geq m$ and the initial concentrations of S and $X(m,n)$ are sufficiently low. Our model similarly suggests that the kinetics of the reverse reaction approaches that of toehold mediated strand displacement when $m \geq n$. It follows then that when $m \approx n$, and the concentrations are below the critical concentrations c_{crit}, both the forward and the reverse reactions are reasonably fast, and the system is in dynamic equilibrium.

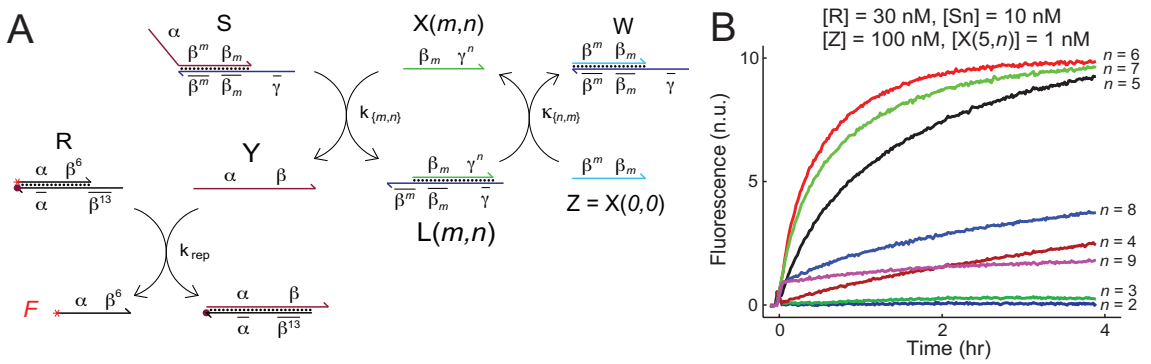


FIG. 3-5: Using toehold exchange to construct a simple catalytic system. **(A)** Schematic of the catalytic system. Strand Z (the same as $X(0,0)$) is present in solution initially with S and R. In the absence of $X(m,n)$, reaction between Z and S to form W and Y is slow (≈ 1 M^{-1} s^{-1}). Both of reverse reactions ($Y + L(m,n) \rightarrow S + X(m,n)$ and $X(m,n) + W \rightarrow Z + L(m,n)$) are present and modeled, with rates constants $k_{\{n,m\}}$ and $\kappa_{\{m,n\}}$, respectively. **(B)** Sample catalysis results. Catalyst $X(5,n)$ was added $t \approx 0$.

We utilize this property to construct a simple reaction network in which $X(m,n)$ acts as catalyst, sustainably speeding up the net reaction $S + Z \rightarrow Y + W$ (Fig. 3-5A). Here $X(0,0)$ is renamed Z because it serves a different role from $X(m,n)$. In the absence of catalyst $X(m,n)$, sequestration of toehold β^m in S should prevent S from reacting with Z; strand Z was shown previously to react with S with a rate constant of about 1 M^{-1} s^{-1}, which is consistent with previous studies on blunt end strand exchange kinetics [64], and is slow enough to be neglected for the purposes of this section. When the two toeholds are of similar intermediate lengths, toehold exchange will ensue rapidly, with $X(m,n)$ reacting with S to yield Y and L. Partially double-stranded L, due to its newly activated toehold β^m, can react quickly with Z to form W and regenerate $X(m,n)$. When excess Z and S are present, each $X(m,n)$ molecule should be able to turn over multiple reactions. Z can therefore be considered a fuel for the catalytic conversion of substrate S to product Y, and W can be considered the waste product.

The net reactions of the system are summarized as:

$$X(m,n) + S \underset{\kappa_{\{n,m\}}}{\overset{k_{\{m,n\}}}{\rightleftharpoons}} Y + L(m,n)$$

$$Z + L(m,n) \underset{k_{\{m,n\}}}{\overset{\kappa_{\{n,m\}}}{\rightleftharpoons}} X(m,n) + W$$

$$Y + R \overset{k_{rep}}{\rightarrow} F$$

Previously, we used $k_{\{m,n\}}$ as shorthand notation for $k_{(\beta^m,\beta_m,\gamma^n)}$. Here, we also need to consider the BM rate constant $k_{(\gamma^n,\beta_m,\beta^m)}$ of the reaction between Z and $L(m,n)$, which we denote as $\kappa_{\{n,m\}}$. In general, $k_{\{m,n\}} \neq \kappa_{\{m,n\}}$, because different sequences are used for the initiation and dissociation toeholds. Unlike the values of $k_{\{m,n\}}$, we do not have any experimental verification on the accuracy of the predicted values of $\kappa_{\{m,n\}}$. The values of $\kappa_{\{m,n\}}$ were evaluated using the calculated binding energies and assuming $k_{f2} = 3 \cdot 10^6$ M^{-1} s^{-1}.

The efficacy of $X(m,n)$ as a catalyst depends on both the relative and the absolute values of m and n. Fig. 3-5B shows the catalytic activities of the set of inputs $X(5,n)$, with n ranging between 2 and 9. When the value n is 2, 3, or 4, the $X(m,n) + S$ reaction is the rate-limiting step of the catalytic cycle. The production rate of F is consistently slow, with $n = 2$ being slowest due to having the slowest forward reaction. When $n = 8$ or 9, the $Z + L(m,n)$ reaction is the rate-limiting step; catalyst $X(m,n)$ is not released rapidly from L by reaction with Z. Consequently, the first turnover is fast, while all subsequent ones are slow, with $n = 9$ yielding the faster initial turnover and slower subsequent speed. When $n = 5$, 6, or 7, both the binding and the release of $X(m,n)$ are fast, reflected in the data by the fast catalytic production of Y: In Fig. 3-5B, the reactions with $X(5,5)$, $X(5,6)$, and $X(5,7)$ approach completion despite substrate S being at a 10-fold higher concentration than the catalyst.

In these catalytic experiments, the initial concentration of Z was 10-fold higher than that of S; this was done to minimize the slowdown as $[Z]$ decreases through the course of the reactions. One side effect of this concentration disparity is to slightly favor the catalytic activity of strands with shorter dissociation toeholds ($m < n$): The reaction between Z and L has higher rate (as opposed to rate constant), and thus less likely to be the rate-limiting step.

The final normalized fluorescence values at $t = 4$ hr, which we call the total catalytic activities, are plotted in Fig. 3-6A for experiments with m =4, 5, 6, and 7. Because $[X] = 1$ nM for these experiments, the total catalytic activities (in nM) corresponds to the turnover (number of reactions catalyzed on average by each catalyst molecule during the

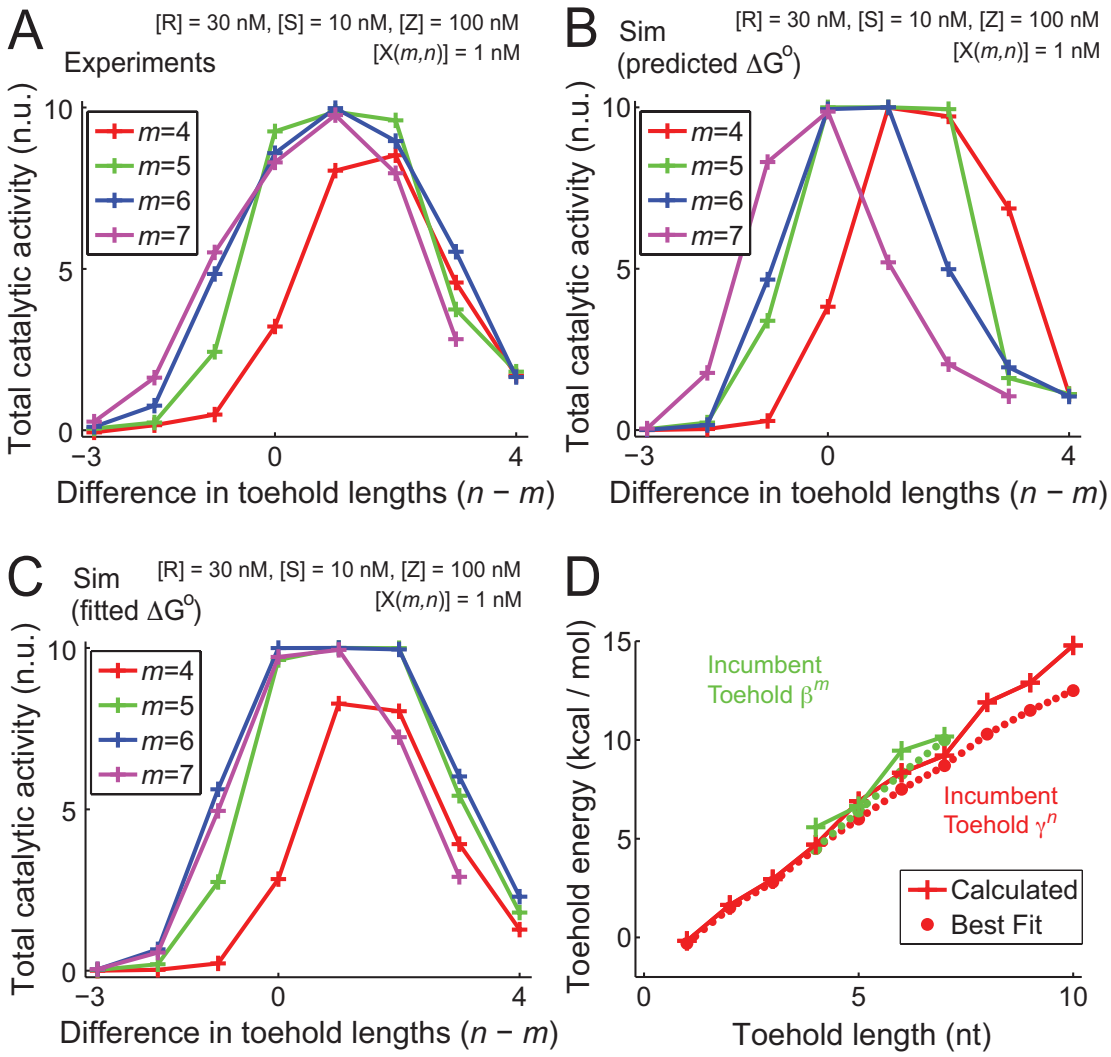


FIG. 3-6: Characterization of catalysis based on toehold exchange. **(A)** Experimental results on the total catalytic turnover over 4 hours. Catalytic activity is greatest when the length of the invading toehold (n) is slightly larger than that of the incumbent toehold (m). **(B)** Simulation results using default energy parameters. In all simulations, $k_f = 3{\cdot}10^6$ M^{-1} s^{-1}, and $k_b = 1.0{\cdot}(\frac{b}{b-m})^2$ M^{-1} s^{-1}. Results are qualitatively similar to experimental data, but show more variation for $n > m$ than was experimentally observed. **(C)** Simulation results using fitted energy parameters. Energy parameters were fitted by hand to to yield good quantitative fitting to the experimental data. **(D)** Differences between energy parameters used in parts (B) and (C).

course of the four hour experiments). The total catalytic activities predicted by the three-step model using the default energy parameters are shown in Fig. 3-6B. For the modeling in this section, the full three-step models of both the $S + X(m,n) \rightleftharpoons L(m,n) + Y$ and the $L(m,n) + Z \rightleftharpoons X(m,n) + W$ reactions were simulated. The reporter reaction $Y + R \rightarrow F$ was modeled as a bimolecular reaction with the previously measured rate constant because

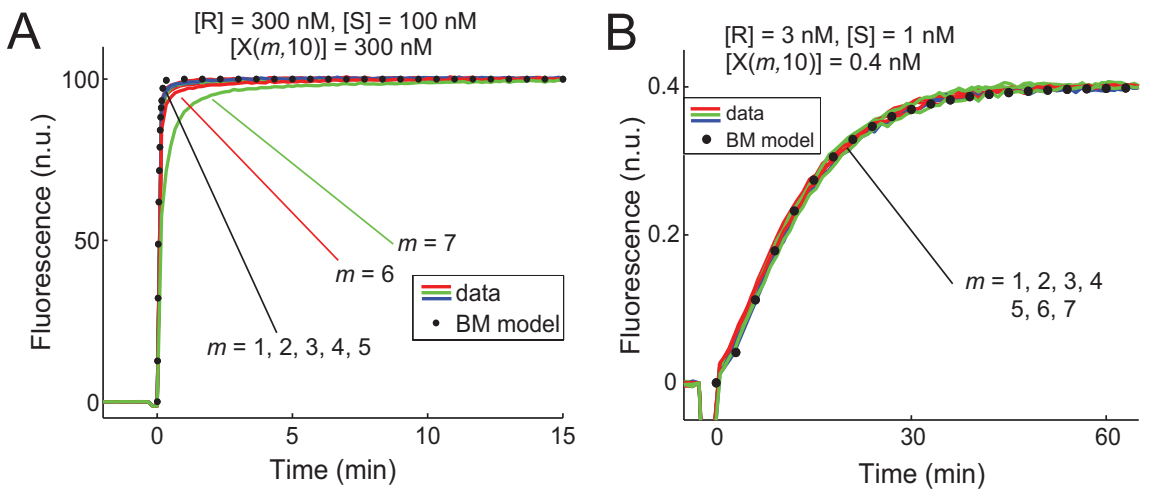


FIG. 3-7: Limitations of models. **(A)** Toehold exchange reactions run at high concentration. The black dotted line shows the predicted behavior for all m under the BM, and predicts reaction kinetics significantly faster than experimentally measured for $m = 6$, 7, and 8. The data and three-step-model results for $m = 6$ differ significantly from those of $m = 7$. The energies used to calculate the reaction rates are those fitted in Fig. 3-6D. **(B)** Toehold exchange reactions run at low concentration. The full three-step model and the BM rate constants produce nearly identical results for all of the strands $X(m, 10)$ (shown by the black dotted line).

the k_f and k_b for that reaction have not been characterized.

The simulation results agreed only qualitatively with experimental data; the discrepancy between the model predictions and results can likely be attributed to inaccuracies in the calculated binding energy of the toeholds, due to the uncertainties discussed in the Materials and Methods section. The three-step model is sensitive to small changes in the relative binding energies of the toeholds because of the exponential role that energy plays in determining rate constants.

To test this hypothesis, binding energies for toeholds were individually fitted to the catalytic data; the best-fit results are shown in Fig. 3-6C, and show significantly better agreement with experimental data. Although there were many free parameters in this fit (10 for the invading toehold energies and 4 for the incumbent toehold energies), the fitted values were all relatively close to predicted (Fig. 3-6D). This provides evidence in support of both the three-step model of toehold exchange, as well as the binding energy models.

Limitations of the BM rate constant.

The BM rate constant is a reasonably accurate predictor of the kinetics of toehold

exchange only when $[X(m,n)]$ and $[S]$ are always below a critical concentration (see Text 3-S2), shown earlier as equation (3):

$$c_{crit} \;=\; \frac{0.1}{k_f} \cdot \frac{k_{r(\gamma^n)}k_{r(\beta^m)} + k_{r(\gamma^n)}k_b + k_{r(\beta^m)}k_b}{k_b + k_{r(\beta^m)}}$$

The critical concentration decreases as the binding energies of the toeholds β^m and γ^n increase. For the systems that we investigated in this paper, the lowest critical concentration occurs in the case of $(m,n) \;=\; (7,10)$, at the value 3.2 nM using the value of $\Delta G^\circ(\beta^7) = -10.2$ kcal/mol fitted in the previous section).

We experimentally test the accuracy of our analysis by comparing toehold exchange kinetics using inputs $X(m,10)$ at high and low concentrations (Fig. 3-7). At low concentrations of $[S] = 1$ nM and $[X(m,10)] = 0.4$ nM, the kinetics of all trajectories are well characterized by the BM rate constants (which are nearly identical for all $m \leq 7$). In contrast, at high concentrations of $[S] = 100$ nM and $[X(m,10)] = 300$ nM, kinetics differ significantly from those predicted by the BM rate constants (black dotted line) for $m = 7$.

Discussion.

By characterizing and quantitatively modeling toehold exchange, this work serves as a "user's manual" for using toehold exchange as a modular design component for hybridization-based dynamical DNA systems. The expressions for the BM rate constant (2) and the critical concentration (3) can be evaluated given the sequences of the nucleic acid strands involved. Fig. 3-8 shows an even simpler method of estimating these two values based on only the lengths of toeholds involved. These estimates assume average strength toeholds, and are derived in text 3-S5. A Matlab script for computing these values is shown in text 3-S6.

Despite the success of the three-step model, some aspects are still imperfectly understood or obviously inaccurate: First, the reasons for dependence of the DNA strand association rate constant (k_f) on the sequence of the toehold domain are not well-understood, differing by a factor of 20 from the fastest sequence to the slowest. The hybridization rate constant k_f

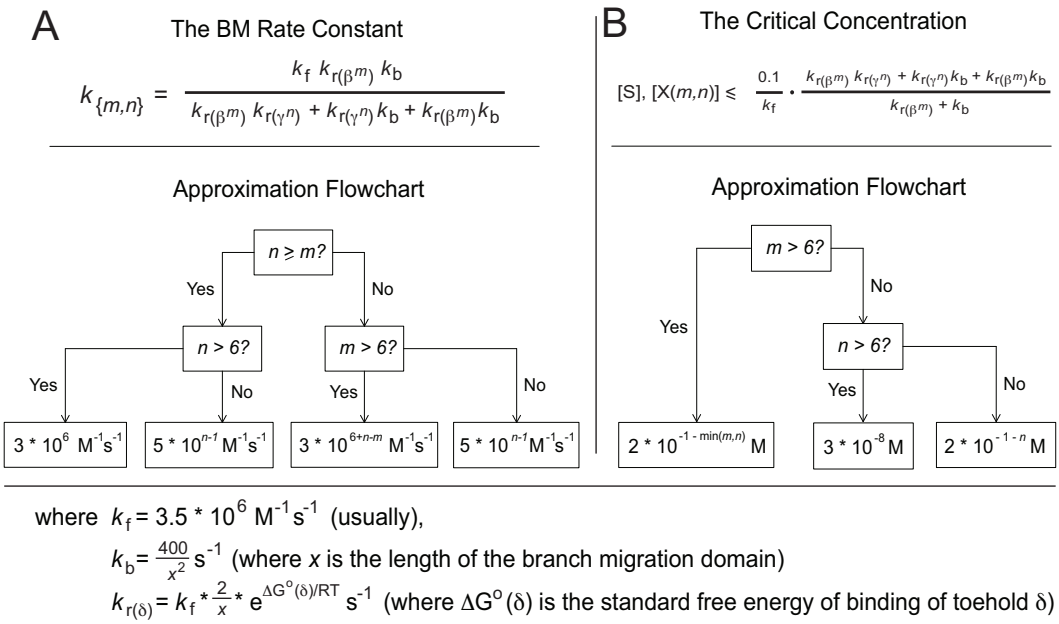


FIG. 3-8: **(A)** Flowchart for order of magnitude estimate of $k_{\{m,n\}}$. This flowchart assumes domain sequences with average binding strength and no secondary structure, and approximates the rate constant at 25 °C and 11.5 mM Mg^{2+}. **(B)** Flowchart for estimating the critical concentration, below which the BM rate constant ($k_{\{m,n\}}$) accurately predicts kinetics.

is thought to depend on hybridization nucleation (i.e. the rate of formation of the first few base pairs) [32, 60, 69]. However, the sequence dependence of k_f is not yet quantitatively predictable, even for secondary structure-free strands.

Second, branch migration is modeled as a single phenomenological rate (k_b) with no intermediate states. This is not only physically inaccurate, but also difficult to predict from first principles. Previous studies have modeled three-stranded branch migration as a random walk process, with time scale $N^2 t_s$, where N is the length of the branch migration domain and t_s is the time scale of an elementary step [42, 43]. For our toehold mediated strand displacement experiments, the length of the branch migration domain β_0 is 20 nt. Naively, our fitted value of $k_b = 1.0\ s^{-1}$ implies an elementary step timescale of $\frac{1}{1.0\ s^{-1}} \cdot \frac{1}{20^2\ \text{steps}} =$ 2.5 ms / step. This value is significantly higher than previously reported results on the rate branch migration of genomic-length DNA (between 10 and 100 μs) [42–44].

However, it is likely that our phenomenological k_b incorporates not only the time needed for branch migration, but also the probability of displacement failure: After $X(m,n)$ and S bind to each other by toehold γ^n, branch migration by the β_m domain ensues. However,

branch migration is inherently a random walk process, and sometimes the junction moves backwards. A significant fraction that start the branch migration process will return to the initial position where $X(m,n)$ is bound to S by only γ^n. In this state, $X(m,n)$ may dissociate before re-starting branch migration. The probability of this occurrence can be estimated mathematically: when n is small, $k_{r(\gamma^n)} >> k_b$, and the probability of dissociation before re-starting branch migration is large and can be approximated as 1. The probability that an unbiased random walk starting at +1 will reach +20 before reaching 0 is simply $\frac{1}{20}$ [70]. Consequently, the actual rate constant of branch migration would then be 20 s^{-1}, corresponding to an elementary step timescale of 125 μs / step, which is much more consistent with previous reported values.

Furthermore, similarly to k_f, the value of k_b may be sequence-dependent; varying the sequence of the branch migration region to discover the range of k_b will be helpful in further understanding the conditions under which our model is accurate. Four-stranded branch migration has been observed to be sequence dependent [68]; three-stranded branch migration can be expected to be more so because of the predominant role that secondary structure can play in single-stranded DNA [54, 68].

Finally, it is hoped that this model is applicable across a range of temperatures and salt concentrations. For example, temperature and salt corrections to the binding thermodynamics of DNA are well-understood [21, 23, 25], and the values of $k_{r(\beta^m)}$ and $k_{r(\gamma^n)}$ can be calculated for different conditions. When the effects of these conditions on k_b are known, it should be possible to predict accurate BM rate constants $k_{(\beta^m,\beta_m,\gamma^n)}$ across a wide range of sequences and conditions without performing any additional experiments (for molecules without significant unwanted secondary structure).

We expect that the same biophysical principles used here to estimate the rate constants of DNA strand displacement are also applicable for understanding toehold exchange within RNA and synthetic nucleic acid analogs (such as LNA and PNA). However, because these other nucleic acid systems generally have stronger Watson-Crick binding interactions and less specificity (for example, G-U wobbles in RNA are significantly more thermodynamically favorable than G-T wobbles in DNA), unwanted secondary structure may be more difficult

to avoid. It is unclear whether higher temperatures and/or more stringent sequence design would be sufficient for obtaining the degree of predictability observed here for DNA toehold exchange kinetics.

Further characterization of the toehold exchange process may lead to improved understanding of the biophysics of nucleic acid hybridization, branch migration, and displacement. Rational design is based on reliable and modular components; detailed characterization of underlying mechanisms and principles will ease the construction of superior nucleic acid logic gates, thresholds, and amplification elements. By interfacing these hybridization-based nucleic acid constructions with other nano-scale engineering accomplishments, DNA, RNA, and synthetic nucleic acids can be used for regulation and timing of more general chemical and biological processes. Previous works in which DNA controlled other nanoscale chemistries [33–40] can be expanded to allow finer control based on the programmed kinetics of DNA. Such advances will enable the construction of robust chemical reaction networks that allow embedded control of biology and chemistry, such as monitoring and modulating the concentrations of biologically relevant molecules.

Acknowledgements. We thank Xi Chen for a very careful reading of this paper and useful suggestions regarding the binding energy calculations. We thank Karthik Sarma for useful suggestions regarding rate constant fitting. We thank Niles Pierce, Justin Bois, and Joe Zadeh for discussion on the energetics parameters used by NUPACK. We thank Niles Pierce, Anne Condon, and Victor Beck for many helpful suggestions in the revision of the manuscript. We thank Bernard Yurke, Georg Seelig, and Joseph Schaeffer for insightful discussions. DYZ and EW were supported by NSF grants 0506468, 0622254, 0533064, 0728703, and 0832824. DYZ is supported by the Fannie and John Hertz Foundation.

3-S1. Quasi-steady state (QSS) derivation of the BM rate constant

Validity of QSS

Many systems of chemical reactions obey QSSA in all but the initial moments of the reaction [62]. QSSA treats the rates of change of the intermediates' concentrations ($\frac{d[\mathrm{I}]}{dt}$ and $\frac{d[\mathrm{J}]}{dt}$ in our system) as small enough to be approximated as 0. The validity of QSSA is ensured when the timescale of the overall reaction is slower than the timescale at which I and J reach their quasi-steady state values. For all experiments presented in this paper except for those in Fig. 3-7, the timescale of the overall reaction is at least 15 minutes, while the timescale of intermediate equilibration is estimated to be on the order of 20 seconds.

The value of parameter k_b was fitted to be 1.0 s^{-1}, so I and J equilibrate with each other on a time scale faster than 1 s. The time constant τ of the initial rise of [I] from 0 to its quasi-steady state value $[\mathrm{I}]_{qss}$ is estimated by $\tau \approx \frac{[\mathrm{I}]_{qss}}{k_f[\mathrm{X}(m,n)][\mathrm{S}]}$. In time τ, the concentration of [I] rises to $[\mathrm{I}]_{qss}(1-\frac{1}{e}) \approx 0.6[\mathrm{I}]_{qss}$.

For convenience, define $x = k_{r(\beta^m)}$ and $y = k_{r(\gamma^n)}$. In equation (7) in the next section, the expression for $[\mathrm{I}]_{qss}$ is seen to be $\frac{k_f(k_b+k_{r(\beta^m)})[\mathrm{X}(m,n)][\mathrm{S}]}{k_{r(\gamma^n)}k_{r(\beta^m)}+k_{r(\gamma^n)}k_b+k_{r(\beta^m)}k_b}$.

$$\begin{aligned}
\tau &\approx \frac{[\mathrm{I}]_{qss}}{k_f[\mathrm{X}(m,n)][\mathrm{S}]} \\
&= \frac{k_f(k_b+x)[\mathrm{X}(m,n)][\mathrm{S}]}{xy+xk_b+yk_b} \cdot \frac{1}{k_f[\mathrm{X}(m,n)][\mathrm{S}]} \\
&= \frac{k_b+x}{xy+xk_b+yk_b}
\end{aligned}$$

If $x > k_b$, $\tau = \frac{k_b+x}{xy+xk_b+yk_b} < \frac{2x}{xy+xk_b+yk_b} < \frac{2x}{xk_b} = \frac{2}{k_b}$. Since the value of k_b was numerically fitted to be 1.0 s^{-1}, this corresponds to a time constant of less than 2 seconds.

If $x < k_b$, $\tau = \frac{k_b+x}{xy+xk_b+yk_b} < \frac{2k_b}{xy+xk_b+yk_b} < \frac{2k_b}{xk_b} = \frac{2}{x}$. From our calculated binding energy of $\beta^7 = -10.17$ kcal/mol and $k_f \approx 3 \cdot 10^6$ M^{-1} s^{-1}, $x \equiv k_{r(\beta^7)} \approx 0.106$ s^{-1}, and the time scale is less than $\frac{2}{0.106} \approx 20$ s.

Thus, QSS is valid for the conditions presented in this paper as well as in similar laboratory circumstances.

Derivation of the BM rate constant

$$\mathrm{X}(m,n) + \mathrm{S} \quad \underset{k_{r(\gamma^n)}}{\overset{k_f}{\rightleftharpoons}} \quad \mathrm{I} \quad \underset{k_b}{\overset{k_b}{\rightleftharpoons}} \quad \mathrm{J} \quad \underset{k_f}{\overset{k_{r(\beta^m)}}{\rightleftharpoons}} \quad \mathrm{Y} + \mathrm{L}(m,n)$$

We aim to use the three-step model of toehold exchange to derive forward and reverse BM rate constants:

$$\mathrm{X}(m,n) + \mathrm{S} \quad \underset{k_{(\gamma^n,\beta_m,\beta^m)}}{\overset{k_{(\beta^m,\beta_m,\gamma^n)}}{\rightleftharpoons}} \quad \mathrm{Y} + \mathrm{L}(m,n)$$

$$\frac{d[\mathrm{Y}]}{dt} = \frac{d[\mathrm{L}(m,n)]}{dt} = k_{(\beta^m,\beta_m,\gamma^n)}[\mathrm{X}(m,n)][\mathrm{S}] \tag{10}$$

$$\frac{d[\mathrm{X}(m,n)]}{dt} = \frac{d[\mathrm{S}]}{dt} = k_{(\gamma^n,\beta_m,\beta^m)}[\mathrm{Y}][\mathrm{L}(m,n)] \tag{11}$$

To derive the expression for $k_{(\beta^m,\beta_m,\gamma^n)}$, we use QSS to analyze the production rate of Y in the absence of the reverse reaction. We first set up the steady state conditions on the intermediates I and J:

$$\frac{d[\mathrm{I}]}{dt} = k_f[\mathrm{X}(m,n)][\mathrm{S}] + k_b[\mathrm{J}] - k_b[\mathrm{I}] - k_{r(\gamma^n)}[\mathrm{I}] \approx 0 \tag{12}$$

$$\frac{d[\mathrm{J}]}{dt} = k_b[\mathrm{I}] + k_f[\mathrm{Y}][\mathrm{L}(m,n)] - k_b[\mathrm{J}] - k_{r(\beta^m)}[\mathrm{J}] \approx 0 \tag{13}$$

In assuming no reverse reaction, we remove the $k_f[\mathrm{Y}][\mathrm{L}(m,n)]$ term in the expression for $\frac{d[\mathrm{J}]}{dt}$. Experimentally, this is achieved by quickly removing Y from the system (using the fluorescence reporter complex), so that the reverse reaction in negligible. Rearranging the simplified equation (13),

$$[\mathrm{J}] = \frac{k_b[\mathrm{I}]}{k_b + k_{r(\beta^m)}} \tag{14}$$

Substituting this expression for [J] back into equation (12) and rearranging,

$$k_f[\mathrm{X}(m,n)][\mathrm{S}] = [\mathrm{I}](\frac{k_{r(\gamma^n)}k_{r(\beta^m)} + k_{r(\gamma^n)}k_b + k_{r(\beta^m)}k_b}{k_b + k_{r(\beta^m)}}) \quad (15)$$

$$[\mathrm{I}] = \frac{k_f(k_b + k_{r(\beta^m)})[\mathrm{X}(m,n)][\mathrm{S}]}{k_{r(\gamma^n)}k_{r(\beta^m)} + k_{r(\gamma^n)}k_b + k_{r(\beta^m)}k_b} \quad (16)$$

$$[\mathrm{J}] = \frac{k_f k_b[\mathrm{X}(m,n)][\mathrm{S}]}{k_{r(\gamma^n)}k_{r(\beta^m)} + k_{r(\gamma^n)}k_b + k_{r(\beta^m)}k_b} \quad (17)$$

Finally, the rate of production of Y is calculated:

$$\begin{aligned}\frac{d[Y]}{dt} &= k_{r(\beta^m)}[J] \\ &= \frac{k_{r(\beta^m)}k_f k_b}{k_{r(\gamma^n)}k_{r(\beta^m)} + k_{r(\gamma^n)}k_b + k_{r(\beta^m)}k_b}[X(m,n)][S]\end{aligned}$$

This BM rate constant expression holds true even when the reverse reaction is considered; see Addendum section 3-S7.

3-S2. Calculations of the critical concentration for the accuracy of the BM rate constant

In the BM, every species that is not product (Y or L(m,n)) is still a reactant; thus, $[\mathrm{X}(m,n)]_{BM} = [\mathrm{X}(m,n)]_0 - [\mathrm{Y}]$ and $[\mathrm{S}]_{BM} = [\mathrm{S}]_0 - [\mathrm{Y}]$, where $[\mathrm{X}(m,n)]_0$ and $[\mathrm{S}]_0$ denote the initial concentrations of X(m,n) and S, and [Y] denotes the measured amount of product Y. In the three-step model $[\mathrm{X}(m,n)] = [\mathrm{X}(m,n)]_0 - [\mathrm{Y}] - [\mathrm{I}] - [\mathrm{J}]$ and $[\mathrm{S}] = [\mathrm{S}]_0 - [\mathrm{Y}] - [\mathrm{I}] - [\mathrm{J}]$.

When conditions are such that [I] and [J] are low (i.e. $[\mathrm{X}(m,n)]_{BM} = [\mathrm{X}(m,n)] + [\mathrm{I}] + [\mathrm{J}] \approx [\mathrm{X}(m,n)]$ and $[\mathrm{S}]_{BM} = [\mathrm{S}] + [\mathrm{I}] + [\mathrm{J}] \approx [\mathrm{S}]$), the kinetics of toehold exchange is well-modeled by a bimolecular reaction with the following BM rate constant:

$$k_{(\beta^m,\beta_m,\gamma^n)} = \frac{k_{r(\beta^m)}k_f k_b}{k_{r(\gamma^n)}k_{r(\beta^m)} + k_{r(\gamma^n)}k_b + k_{r(\beta^m)}k_b} \quad (18)$$

In contrast, when the concentrations of I and J are high, the above BM rate constant will overestimate the kinetics of the toehold exchange reaction.

To evaluate the conditions under which the BM rate constant grossly overestimates the

kinetics of toehold exchange, we analyze the concentration of [I]: We arbitrarily define the condition $[\mathrm{X}(m,n)] \approx [\mathrm{X}(m,n)]_{BM} = [\mathrm{X}(m,n)]+[\mathrm{I}]+[\mathrm{J}]$ to be satisfied when $\frac{[\mathrm{I}]}{[\mathrm{X}(m,n)]} \leq 0.1$. Because $[\mathrm{J}] = [\mathrm{I}]\frac{k_b}{k_b+k_{r(\beta^m)}} < [\mathrm{I}]$, $[\mathrm{X}(m,n)]_{BM} < 1.2 \cdot [\mathrm{X}(m,n)]$. Similarly, $[\mathrm{S}] \approx [\mathrm{S}]_{BM} = [\mathrm{S}]+[\mathrm{I}]+[\mathrm{J}] < 1.2 \cdot [\mathrm{S}]$ when $\frac{[\mathrm{I}]}{[\mathrm{S}]} \leq 0.1$. The expression for the quasi-steady steady production rate of Y depends on product of $[\mathrm{X}(m,n)]$ and [S]:

$$\begin{aligned} \frac{d[\mathrm{Y}]}{dt} &= k_{(\beta^m,\beta_m,\gamma^n)}[\mathrm{X}(m,n)][\mathrm{S}] \\ &> k_{(\beta^m,\beta_m,\gamma^n)}(\frac{1}{1.2}[\mathrm{X}(m,n)]_{BM})(\frac{1}{1.2}[\mathrm{S}]_{BM}) \end{aligned}$$

$$k_{(\beta^m,\beta_m,\gamma^n)}[\mathrm{X}(m,n)]_{BM}[\mathrm{S}]_{BM} < 1.44\frac{d[\mathrm{Y}]}{dt}$$

Thus when $\frac{[\mathrm{I}]}{[\mathrm{X}(m,n)]}, \frac{[\mathrm{I}]}{[\mathrm{S}]} \leq 0.1$, the rate of production of Y predicted by the derived BM rate constant is overestimated by no more than 44%. Rearranging (4), we solve for the critical concentration of S below which $\frac{[\mathrm{I}]}{[\mathrm{X}(m,n)]} \leq 0.1$:

$$[\mathrm{S}] \leq \frac{0.1}{k_f} \cdot \frac{k_{r(\gamma^n)}k_{r(\beta^m)} + k_{r(\gamma^n)}k_b + k_{r(\beta^m)}k_b}{k_b + k_{r(\beta^m)}} \tag{19}$$

Similar derivation on the condition $\frac{[\mathrm{I}]}{[\mathrm{S}]} \leq 0.1$ leads to the same critical concentration for $\mathrm{X}(m,n)$. The critical concentration varies monotonically with the figure of merit; for example, the BM predicted kinetics are off by no more than an order of magnitude when $[\mathrm{X}(m,n)]$ and [S] are below $\frac{1}{k_f} \cdot \frac{k_{r(\gamma^n)}k_{r(\beta^m)}+k_{r(\gamma^n)}k_b+k_{r(\beta^m)}k_b}{k_b+k_{r(\beta^m)}}$ (when $\frac{[\mathrm{I}]}{[\mathrm{X}(m,n)]}, \frac{[\mathrm{I}]}{[\mathrm{S}]} \leq 1$).

3-S3. Calculation of Toehold Binding Energies

The binding energies of the toehold are calculated as:

$$\begin{aligned} \Delta G^\circ(\gamma^n) &= \Delta G^\circ(\mathrm{I}(0,n)) - \Delta G^\circ(\mathrm{S}) \\ \Delta G^\circ(\beta^m) &= \Delta G^\circ(\mathrm{J}(m,0)) - \Delta G^\circ(\mathrm{L}(m,0)) \end{aligned}$$

Default calculation method

Here, we show a step-by-step method for calculating ΔG° of two complexes, S, and I(0,3), which is in turn used to infer the binding energy of the toehold γ^3. Calculation of ΔG° values for other complexes are analogous.

$\Delta G^\circ(\mathbf{S})$

We start by calculating the ΔG° of S:

$$\Delta G^\circ(\mathrm{S}) \;=\; 1 * \Delta G^\circ_{\mathrm{init}} + \Delta G^\circ_{\alpha\ \mathrm{dangle}} + \Delta G^\circ_{\gamma\ \mathrm{dangle}} + \Delta G^\circ_{\mathrm{stacks}}$$

For each of the ΔG° terms, $\Delta G^\circ = \Delta H^\circ - T\Delta S^\circ$.

Hybridization initiation. The $\Delta G^\circ_{\mathrm{init}}$ refers to the energetic cost of initiating a helix. SantaLucia et al. [21] report that $\Delta H^\circ_{\mathrm{init}} = +0.2$ kcal/mol and $\Delta S^\circ_{\mathrm{init}} = -5.7$ cal/mol·K, leading to $\Delta G^\circ_{\mathrm{init}} = +1.90$ kcal/mol at 25 °C.

Terminal dangles. For S, the terminal dangle on the α domain is a 5' T dangle with a C nearest neighbor. Bommarito et al. [22] report $\Delta H^\circ_{\alpha\ \mathrm{dangle}} = -4.0$ kcal/mol and $\Delta S^\circ_{\alpha\ \mathrm{dangle}} = -10.9$ cal/mol·K, leading to $\Delta G^\circ_{\alpha\ \mathrm{dangle}} = -0.75$ kcal/mol at 25 °C.

For S, the terminal dangle on the γ domain is a 5' A dangle with a C nearest neighbor. Bommarito et al. [22] report $\Delta H^\circ_{\gamma\ \mathrm{dangle}} = -6.3$ kcal/mol and $\Delta S^\circ_{\gamma\ \mathrm{dangle}} = -17.1$ cal/mol K, leading to $\Delta G^\circ_{\gamma\ \mathrm{dangle}} = -1.20$ kcal/mol at 25 °C. For Ss and Sw, the γ dangle would be a 5' G and a 5' T, respectively.

Stacks. There are 20 base pairs in the β domain of S, leading to 19 total stack terms. Summing these using the values reported by SantaLucia et al. [21] yield a total $\Delta H^\circ_{\mathrm{stacks}} = -152.4$ kcal/mol and $\Delta S^\circ_{\mathrm{stacks}} = -409.4$ cal/mol·K, leading to $\Delta G^\circ_{\mathrm{stacks}} = -30.40$ kcal/mol at 25 °C.

Total standard free energy. Summing all the previous terms, $\Delta G^\circ(\mathrm{S}) = -30.45$ kcal/mol.

$\Delta G^\circ(\mathbf{I(0,3)})$

The binding energy of I(0,3) is calculated as:

$$\begin{aligned}\Delta G^\circ(\text{I}(0,3)) \;=\; & 2 * \Delta G^\circ_{\text{init}} + \Delta G^\circ_{\alpha\ \text{dangle}} + \Delta G^\circ_{\gamma\ \text{dangle}} + \Delta G^\circ_{\text{stacks}} \\ & + \Delta G^\circ_{\text{AT termination}} + \Delta G^\circ_{\text{nick}} + \Delta G^\circ_{\text{coaxial dangle}}\end{aligned}$$

Initiation. $\Delta G^\circ_{\text{init}} = +1.90$ kcal/mol as calculated before, but we multiply it by 2 in this calculation because I(0,3) is composed of three strands of DNA.

Terminal dangles. The α dangle is exactly the same, so $\Delta G^\circ_{\alpha\ \text{dangle}} = -0.75$ kcal/mol. The γ dangle in this case is a 5' G with a A nearest neighbor, which is reported to contribute $\Delta H^\circ_{\gamma\ \text{dangle}} = -1.1$ kcal/mol and $\Delta S^\circ_{\gamma\ \text{dangle}} = -1.6$ cal/mol·K, leading to $\Delta G^\circ_{\gamma\ \text{dangle}} = -0.62$ kcal/mol.

Stacks. I(0,3) has 23 total base pairs, and consequently 22 stacks, but one of these stacks is the coaxial stacking term at the nick. Summing the remaining 21 stacks yields $\Delta H^\circ_{\text{stacks}} = -168.4$ kcal/mol and $\Delta S^\circ_{\text{stacks}} = -452.6$ cal/mol·K, leading to $\Delta G^\circ_{\text{stacks}} = -33.53$ kcal/mol.

AT termination. The $\Delta G^\circ_{\text{AT termination}}$ is a special energetic penalty term that is added for helices terminating in an A-T base pair. This term is part of the commonly accepted energy parameters presented by SantaLucia et al. [21]. $\Delta H^\circ_{\text{AT termination}} = +2.2$ kcal/mol and $\Delta S^\circ_{\text{AT termination}} = +6.9$ cal/mol·K, leading to $\Delta G^\circ_{\text{AT termination}} = +0.14$ kcal/mol.

Coaxial stacking. The $\Delta G^\circ_{\text{nick}}$ is the coaxial stacking term at the boundary of the β and γ domains. The bases flanking the nick are 5'-G and T-3'. Protozanova et al. [26] report the ΔG° of this coaxial stack to be -2.04 kcal/mol at 37 °C [26]. Unfortunately, this work did not provide explicit ΔH° and ΔS° values, so the ΔG° value at 25 °C cannot be directly calculated.

Protozanova et al. [26] suggested the approximation that $\Delta S^\circ \approx -25$ cal/mol·K for all coaxial stacks in order to estimate ΔG° at other temperatures. This would imply that $\Delta G^\circ_{25} = \Delta G^\circ_{37} - 0.3$ kcal/mol for all coaxial stacks.

Observing the distribution of ΔS° and ΔH° values for standard stacks [21, 30, 31], we felt that $\Delta S^\circ \approx a\Delta H^\circ$ was a better approximation than $\Delta S^\circ \approx 25$ cal/mol·K, where a is a fitted constant (see Fig. 3-S1). A standard least squares fit yielded $a = 0.0027$ K^{-1}.

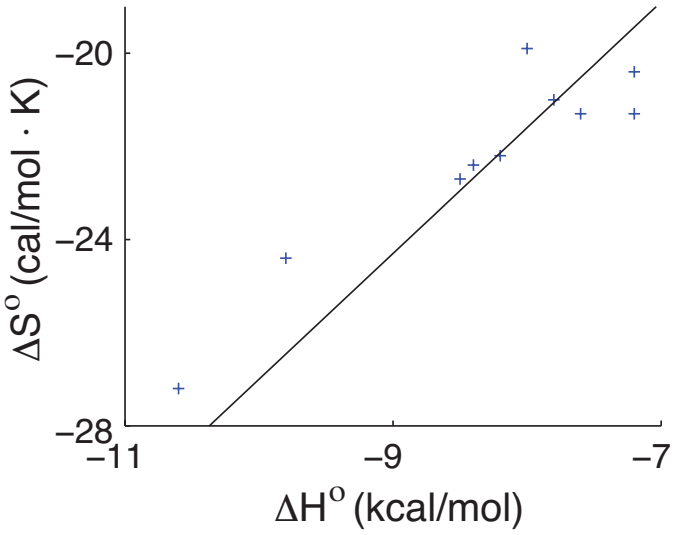


FIG. 3-S1: Distribution of ΔH° and ΔS° values for standard nearest-neighbor stacks, as reported by ref. [21]. The black line indicates the approximation $\Delta S^\circ = 0.0027\text{K}^{-1}\Delta H^\circ$.

Because $\Delta G^\circ = \Delta H^\circ - T\Delta S^\circ$, $\Delta G^\circ_{37} \approx (1-310*0.0027)\Delta H^\circ = 0.163\Delta H^\circ$ and $\Delta G^\circ_{25} \approx (1-298*0.0027)\Delta H^\circ = 0.195\Delta H^\circ$. Dividing the two equations and rearranging, $\Delta G^\circ_{25} = 1.2 \cdot \Delta G^\circ_{37}$. For the coaxial stack in question, $\Delta G^\circ_{\text{nick}} \approx 1.2 \cdot -2.04 = -2.45$ kcal/mol.

Protozanova et al.'s reported values for the ΔG°_{37} of coaxial stacks [26] range from -0.12 kcal/mol to -2.70 kcal/mol. Using our approximation leads to ΔG°_{25} ranging from -0.144 to -3.24 kcal/mol, while using $\Delta S^\circ \approx 25$ cal/mol·K leads to ΔG°_{25} ranging from -0.42 to -3.00 kcal/mol. Thus, the two approximations should differ by no more than 0.3 kcal/mol in all cases.

Coaxial stack dangles. The $\Delta G^\circ_{\text{coaxial dangle}}$ term refers to the energy contribution of the last G on the β domain of input X(m,n), when the latter is bound to S by only the toehold. As explained in the main paper, the value of this parameter has not been characterized, and is assumed to be 0.

Total standard free energy. Summing the terms above, $\Delta G^\circ(\text{I}(0,3)) = -33.41$ kcal/mol. The binding strength of toehold γ^3 is inferred to be $\Delta G^\circ(\text{I}(0,3)) - \Delta G^\circ(\text{S}) = -2.95$ kcal/mol. This value is then rounded to 1 decimal point and are shown in Table 3-2 and Table 3-S1.

Pyshnyi's coaxial stacking terms

For this energetics model, calculations were performed completely analogously to the

default method, except using the coaxial stacking parameters reported by Pyshnyi et al. [27]. Pyshnyi et al. did report ΔH° and ΔS° values for all of their coaxial stacking terms, so ΔG°_{25} was directly calculated for the coaxial stack. The inferred toehold binding energies are shown in column 3 of Table 3-S1.

NUPACK + Protozanova

For energetics calculations using the NUPACK folding software [17], we used the "energy" program in the downloadable NUPACK 2.1 software. This program computes the standard free energy of a particular microstate entered by the user, entered in dot-paren notation. For example, our input file for I(0,3) was the following:

```
CCACATACATCATATTCCCTCATTCAATACCCTACG
CCCTCATTCAATACCCTACGTCT
GAAGTGACATGGAGACGTAGGGTATTGAATGAGGG
1 2 3
.....((((((((((((((((((((+.....((((+.....))))))))))))))))))))))
```

According to the user's manual for NUPACK, calculations using the "dangles = some" option (the recommended default) calculates dangle energies "for each unpaired base flanking a duplex (a base flanking two duplexes contributes only the minimum of the two possible energies)." Thus, the "dangles = some" option does not include any coaxial stacking energetics term, but does include the coaxial stacking dangle term (with energy equivalent to an analogous terminal dangle). We thus manually add our modified Protozanova et al. coaxial stacking terms to the standard free energy of complexes predicted by NUPACK, where applicable.

Additionally, as of this writing, NUPACK calculates the standard free energy of complexes to satisfy the thermodynamic equilibria of the molecules in mole fraction, rather than concentration (in molar). This means that an additional corrective term needs to be manually added for calculating the toehold binding energy.

As an example, consider $\Delta G^\circ(\mathrm{S})$:

$$\begin{aligned} \Delta G^\circ_{NUPACK} &= -RT\ln(K_{NUPACK}) \\ K_{NUPACK} &= \frac{\chi_{\mathrm{S}}}{\chi_{\mathrm{Y}}\chi_{\mathrm{V}}} \end{aligned}$$

where V is the heretofore unnamed bottom strand of the S complex (containing the domains $\bar{\gamma}$ and $\bar{\beta}$. To convert the equilibrium constant K to be expressed in terms of M^{-1}, a correction $RT\ln(c)$ term needs to be added, where c is the total concentration of all species in solution. The total concentration of all species is solution is dominated by the solvent water molecules, which exist at 55 M, implying that the correction is roughly $RT\ln(55) \approx$ 2.38 kcal/mol. Note that this correction is to be multiplied by the $N-1$, where N is the total number of strands in the complex. For the energies listed in column 4 in Table 3-S1, we manually added the correction where applicable.

NUPACK, dangles = all

According to the user's manual for NUPACK, calculations using the "dangles = all" option calculates dangle energies "for each base flanking a duplex regardless of whether it is paired." Thus, the "dangles = all" option includes not only the coaxial stacking dangle term as before, but also approximates the coaxial stack energy as the sum of the two dangles. For example, in I(0,n), the nick is "G / T"; the energetics of this nick is approximated by NUPACK as the sum of that of 5' A dangle with nearest neighbor C, and a 3' C dangle with nearest neighbor A. This method of approximation is not scientifically justified, but rather exists as a placeholder until coaxial stacking energetics are better understood.

As in the previous method, manual corrective terms of 2.38 kcal/mol per extra strand was added to convert ΔG° for mole fractions to ΔG° for molar units. The corrected inferred toehold binding energies are shown in column 5 of Table 3-S1.

Owczarzy's Mg^{2+} correction

Owczarzy et al.'s correction formula [25] modifies ΔS° based on the ΔH° of the structure in 1 M Na^+. The formula depended on the length of the helix involved; for our calculations, we assumed that the nick does not disrupt the helix (e.g. I(0,3) would contain a helix of length 23, rather than 2 helices of lengths 20 and 3). Furthermore, since this study did not characterize the salt corrections for energy contributions from dangles, we used Bommarito et al.'s dangle values (for 1 M Na^+) [22].

Mathematically, the formula we used was:

$$\begin{aligned}\Delta H^{\circ}_{\mathrm{Mg^{2+}}} &= \Delta H^{\circ}_{\mathrm{Na^{+}}} \\ \Delta S^{\circ}_{\mathrm{Mg^{2+}}} &= \Delta S^{\circ}_{\mathrm{Na^{+}}} + \Delta H^{\circ}_{stacks,\mathrm{Na^{+}}} \cdot (a + bx + f_{\mathrm{GC}}(c + dx) + \frac{e + fx + gx^2}{2(N_{\mathrm{bp}} - 1)})\end{aligned}$$

where $x = \ln[\mathrm{Mg^{2+}}] = -4.465$, and f_{GC} is the fraction of the bases in the helix that are purines. Parameters $a = 3.92 \cdot 10^{-5}$ K^{-1}, $b = -9.11 \cdot 10^{-5}$ K^{-1}, $c = 6.26 \cdot 10^{-5}$ K^{-1}, $d = 1.42 \cdot 10^{-5}$ K^{-1}, $e = -4.82 \cdot 10^{-4}$ K^{-1}, $f = 5.25 \cdot 10^{-4}$ K^{-1}, and $g = 8.31 \cdot 10^{-5}$ K^{-1} are fitted values reported by Owczarzy et al. [25]. The toehold binding energies calculated using this model are shown in column 6 of Table 3-S1.

Toehold	Default	Pyshnyi coaxial	NUPACK, dangles=some	NUPACK, dangles=all	Owczarzy Mg^{2+}
γ^0	+1.9	+1.9	+1.9	+2.5	+1.9
γ^1	+0.2	+0.2	+0.5	+1.5	+0.3
γ^2	-1.7	-1.6	-1.2	-0.3	-1.4
γ^3	-3.0	-2.9	-2.6	-1.6	-2.5
γ^4	-4.7	-4.7	-4.3	-3.4	-4.1
γ^5	-6.9	-6.9	-6.5	-5.5	-6.1
γ^6	-8.3	-8.3	-7.9	-6.9	-7.3
γ^7	-9.2	-9.2	-8.8	-7.9	-8.0
γ^8	-11.9	-11.9	-11.4	-10.5	-10.5
γ^9	-12.9	-12.9	-12.5	-11.5	-11.3
γ^{10}	-14.8	-14.8	-14.3	-13.4	-13.0
γ^{15}	-21.8	-21.8	-21.4	-20.4	-19.1
γs^0	+1.9	+1.9	+1.9	+2.5	+1.9
γs^1	-1.1	-1.0	-2.0	+0.2	-1.0
γs^2	-3.2	-3.1	-4.1	-1.9	-2.9
γs^3	-5.0	-5.0	-6.0	-3.8	-4.6
γs^4	-8.0	-7.9	-8.9	-6.7	-7.3
γs^5	-10.3	-10.2	-11.2	-9.0	-9.4
γs^6	-12.1	-12.0	-13.1	-10.9	-11.0
γs^7	-15.1	-15.0	-16.0	-13.8	-13.8
γs^8	-17.3	-17.3	-18.3	-16.1	-15.8
γs^9	-19.2	-19.1	-20.2	-18.0	-17.5
γs^{10}	-21.2	-21.2	-22.2	-20.0	-19.3
γw^0	+1.9	+1.9	+1.9	+2.5	+1.9
γw^1	+0.2	-0.6	-0.4	+1.2	+0.2
γw^2	-0.8	-1.5	-1.4	+0.2	-0.6
γw^3	-2.1	-2.8	-2.6	-1.1	-1.6
γw^4	-3.8	-4.5	-4.4	-2.8	-3.2
γw^5	-4.3	-5.0	-4.9	-3.3	-3.5
γw^6	-5.3	-6.0	-5.9	-4.3	-4.3
γw^7	-7.0	-7.7	-7.6	-6.0	-5.9
γw^8	-7.5	-8.2	-8.1	-6.5	-6.2
γw^9	-8.9	-9.7	-9.6	-8.0	-7.5
γw^{10}	-8.9	-9.6	-9.4	-7.9	-7.2

Table 3-S1: Comparison of toehold binding energies using various methods (in kcal/mol)

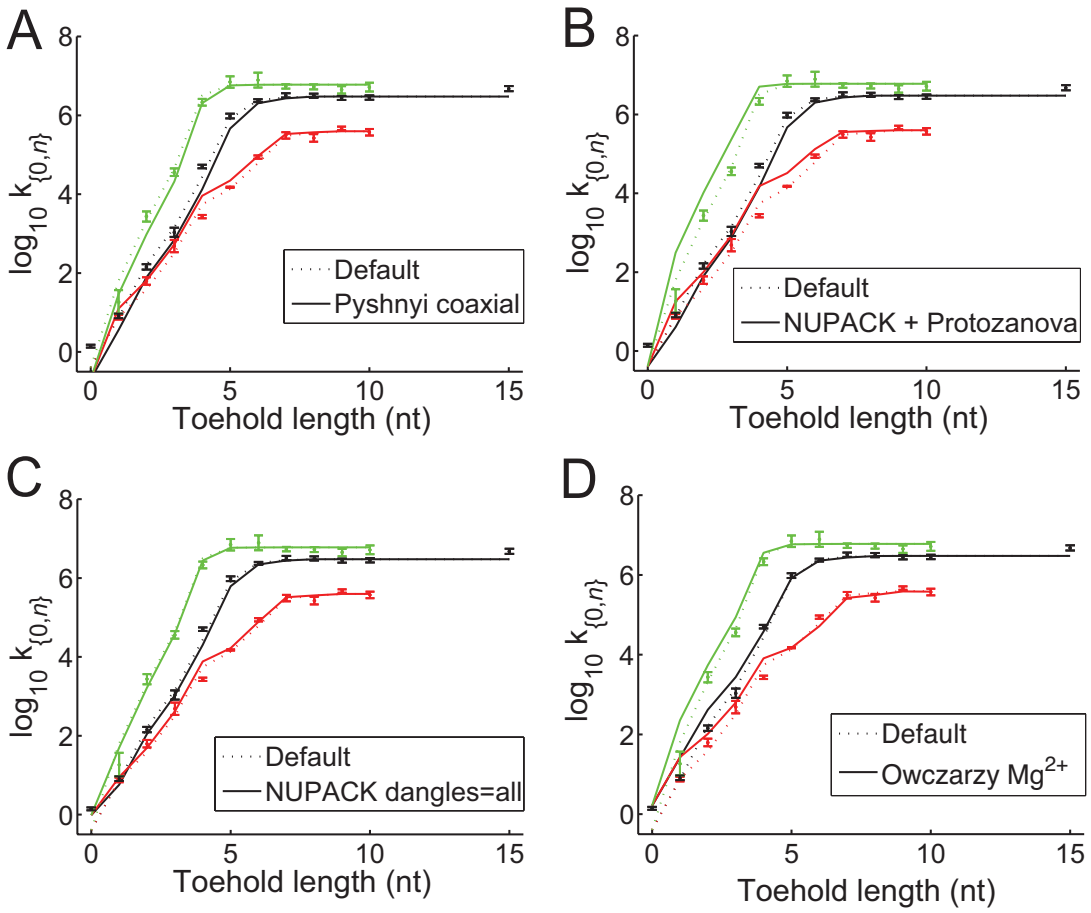


FIG. 3-S2: Comparison of different energy models and their effects on the model-predicted BM rate constant. The dots represent the best-fit rate constants from data in the main paper. The dotted lines show the model-predicted BM rate constants based on the toehold binding energies listed in the main paper. **(A)** The solid lines show the BM rate constants predicted using coaxial stacking terms reported by Pyshnyi et al. [27], but otherwise identical to the default model. For this set of energy parameters, $k_b = 0.5\ \mathrm{s}^{-1}$. **(B)** The solid lines show the BM rate constants predicted using energies predicted by NUPACK [17], using "dangles = some," so that the coaxial stack dangle are calculated as terminal dangles. The modified Protozanova et al. [26] coaxial stacking parameter was manually added, as well as a +2.377 kcal/mol conversion term to convert the standard free energy to molar units (from mole fraction). See Text S3 for details on how NUPACK was used. For this set of energy parameters, $k_b = 1\ \mathrm{s}^{-1}$. **(C)** The solid lines show the BM rate constants predicted using energies predicted by NUPACK, using "dangles = all." The coaxial stack dangles are calculated as terminal dangles, and the coaxial stacking parameter is calculated as two overlapping dangles. A +2.377 kcal/mol conversion term to convert the standard free energy to molar units (from mole fraction). For this set of energy parameters, $k_b = 7\ \mathrm{s}^{-1}$. **(D)** The solid lines show the BM rate constants predicted using toehold binding energies as calculated by the default model, corrected for magnesium concentration using the method by Owczarzy et al. [25]. For this set of energy parameters, $k_b = 4\ \mathrm{s}^{-1}$.

m	n	[S]	[X(m,n)]	[R]
0	0-1	100 nM	2 μM, 4 μM, and 6 μM	300 nM
0	2	100 nM	200 nM, 400 nM, and 600 nM	300 nM
0	3	10 nM	20 nM, 40 nM, and 60 nM	30 nM
0	4	10 nM	2 nM, 4 nM, and 6 nM	30 nM
0	5-10, 15	1 nM	0.2 nM, 0.4 nM, and 0.6 nM	3 nM
1-7	10	1 nM	0.4 nM	3 nM
5-7	9	1 nM	0.4 nM	3 nM
4-7	8	1 nM	0.4 nM	3 nM
4-7	7	1 nM	0.4 nM	3 nM
2-5	6	1 nM	0.4 nM	3 nM
6-7	6	10 nM	4 nM	30 nM
1-4	5	1 nM	0.4 nM	3 nM
5-6	5	10 nM	4 nM	30 nM
7	5	10 nM	40 nM	30 nM
1-7	4	10 nM	40 nM	30 nM
1-3	3	10 nM	40 nM	30 nM
4-5	3	10 nM	400 nM	30 nM
6	3	10 nM	40 nM	30 nM
1-5	2	10 nM	400 nM	30 nM
1-3	1	10 nM	4 μM	30 nM
4	1	10 nM	400 nM	30 nM

Table 3-S2: Concentrations used for toehold exchange reactions using X(m,n)

m	n	[Sw]	[Xw(m,n)]	[R]
0	1	10 nM	2 μM, 4 μM, and 6 μM	30 nM
0	2	10 nM	200 nM, 400 nM, and 600 nM	30 nM
0	3-5	10 nM	20 nM, 40 nM, and 60 nM	30 nM
0	6	1 nM	2 nM, 4 nM, and 6 nM	3 nM
0	7-10	1 nM	0.2 nM, 0.4 nM, and 0.6 nM	3 nM

Table 3-S3: Concentrations used for toehold mediated strand displacement using Xw(m,n)

m	n	[Ss]	[Xs(m,n)]	[R]
0	1	10 nM	2 μM, 4 μM, and 6 μM	30 nM
0	2	10 nM	20 nM, 40 nM, and 60 nM	30 nM
0	3	10 nM	2 nM, 4 nM, and 6 nM	30 nM
0	4-10	1 nM	0.2 nM, 0.4 nM, and 0.6 nM	3 nM

Table 3-S4: Concentrations used for toehold mediated strand displacement using Xs(m,n)

3-S4. Sample Matlab code used for fitting rate constants.

The main program for fitting the BM rate constant of toehold exchange using input X(4,8) is as follows:

```
k0 = log(1E6);
scale0 = log(4e13);

[k, fval] = fminunc(@TE_4_8, [k0, scale0]);
```

The variable $k0$ shows an initial "guess" of the rate constant (set to 10^6 M^{-1} s^{-1} here), while $scale0$ shows an initial guess of the scaling constant (set to $4 \cdot 10^{13}$ counts per mole of unquenched fluorophore). Both variables are fitted during runtime.

The error function TE_4_8 for this is as follows:

```
function err_func = TE_4_8(input)

data = load('/Users/daveyzhang/Desktop/work/expt/Fluorescence/20080229/004m.txt');

k = exp(input(1));
scalingconst = exp(input(2));

err_func = 0;
options = odeset('RelTol', 1e-4, 'AbsTol', 1e-30);

datasize = size(data, 1);

%(4,8) data
t = data(6:datasize,1)-300;
y0 = [1e-9, 4e-10, 0, 3e-9, 0];
[t, y2] = ode23s(@rdy, t, [k, y0], options);

ye = y2(:,6) * scalingconst + data(5,2);
for i = 7:size(data, 1)
    err_func = err_func + (ye(i-5) - data(i,2))^2/ye(i-5);
end
```

The variable $y0$ sets the initial conditions of the simulation, and the Matlab function $ode23s$ is used to simulated the reactions defined in the function rdy. Note that we use

*ode*23*s* rather than *ode*45 because the system is "stiff," containing reactions with very different time scales.

The *rdy* function is as follows:

```
function dy = rdy(t, y)
%S + X -> OB
%OB + OF_OQ -> OF
krep = 1.30e6; %fitted earlier
dy = zeros(6,1);
dy(2) = -y(1) * y(2) * y(3);
dy(3) = -y(1) * y(2) * y(3);
dy(4) = y(1) * y(2) * y(3) - krep * y(4) * y(5);
dy(5) = -krep * y(4) * y(5);
dy(6) = krep * y(4) * y(5);
```

The parameter *krep* denotes the rate constant of the reaction between the output product Y and the reporter complex R (previously fitted to be $1.3 \cdot 10^6$ M^{-1} s^{-1}.

3-S5. Approximation of the BM rate constants and critical concentrations.

In Fig. 3-8 of the main text, we show a simplified approximation flowchart for estimating the BM rate constant and the critical concentration below which the BM rate constant is a valid predictor of kinetics. Here, we justify the results in those figures.

Approximating the BM rate constant

Recall equation (7), the expression for the BM rate constant:

$$\begin{aligned} k_{\{m,n\}} &\equiv k_{(\beta^m,\beta_m,\gamma^n)} \\ &= \frac{k_f k_{r(\beta^m)} k_b}{k_{r(\gamma^n)} k_{r(\beta^m)} + k_{r(\gamma^n)} k_b + k_{r(\beta^m)} k_b} \end{aligned}$$

At room temperature of 25 °C, with 11.5 mM Mg^{2+}, and average distribution of base pairs for toeholds β^m and γ^n, each base contributes on average approximately 1.4 kcal/mol to

the binding energy of the toehold. Numerical substitution yields:

$$\begin{aligned}
\Delta G^{\circ}(\beta^m) &\approx 1.9 - 1.4m[\text{kcal/mol}] \\
k_{r(\beta^m)} &= k_f \frac{2}{b-m} e^{\Delta G^{\circ}(\beta^m)/RT} \\
&= 3 \cdot 10^6 \frac{2}{20-m} (e^{-2.36})^{(m-1.35)} [\text{M}^{-1}\ \text{s}^{-1}] \\
&= 3 \cdot 10^6 \frac{2}{20-m} (10.6)^{1.35-m} [\text{M}^{-1}\ \text{s}^{-1}] \\
&\approx 6 \cdot 10^{6-m} [\text{M}^{-1}\ \text{s}^{-1}]
\end{aligned}$$

Similarly, $k_{r(\gamma^n)} \approx 6 \cdot 10^{6-n}[\text{M}^{-1}\ \text{s}^{-1}]$. Compared to the empirically measured value of $k_b = 1.0\ \text{s}^{-1}$, $k_{r(\beta^m)}$ and $k_{r(\gamma^n)}$ are smaller than k_b when m and n are greater than 6, respectively.

For situations where $n \geq m$, $k_{r(\beta^m)} \geq k_{r(\gamma^n)}$, and the expression for $k_{\{m,n\}}$ can be approximated as:

$$k_{\{m,n\}} \approx \frac{k_f k_{r(\beta^m)} k_b}{(k_{r(\gamma^n)} + k_b) k_{r(\beta^m)}} = \frac{k_f k_b}{k_{r(\gamma^n)} + k_b}$$

When $n > 6$, $k_{r(\gamma^n)} < k_b$, and the expression for $k_{\{m,n\}}$ is approximated as $k_f \approx 3 \cdot 10^6\ \text{M}^{-1}\ \text{s}^{-1}$. When $n \leq 6$, $k_{r(\gamma^n)} > k_b$, and the expression for $k_{\{m,n\}}$ is approximated as $\frac{k_f k_b}{k_{r(\gamma^n)}} = 5 \cdot 10^{n-1}\ \text{M}^{-1}\ \text{s}^{-1}$.

For situations where $n < m$, $k_{r(\beta^m)} < k_{r(\gamma^n)}$, and the expression for $k_{\{m,n\}}$ can be approximated as:

$$k_{\{m,n\}} \approx \frac{k_f k_{r(\beta^m)} k_b}{(k_{r(\beta^m)} + k_b) k_{r(\gamma^n)}}$$

When $m > 6$, $k_{r(\beta^m)} < k_b$, and the expression for $k_{\{m,n\}}$ can be approximated as $\frac{k_f k_{r(\beta^m)}}{k_{r(\gamma^n)}} = k_f \cdot 10^{n-m} \approx 3 \cdot 10^{6+n-m}\ \text{M}^{-1}\ \text{s}^{-1}$. When $m \leq 6$, $k_{r(\beta^m)} > k_b$, and the expression for $k_{\{m,n\}}$ can be approximated as $\frac{k_f k_b}{k_{r(\gamma^n)}} = 5 \cdot 10^{n-1}\ \text{M}^{-1}\ \text{s}^{-1}$.

Approximating the critical concentration

For estimating the critical concentrations below which the BM rate constant accurately predicts kinetics, recall the expression for the critical concentration:

$$[\mathrm{X}(m,n)],[\mathrm{S}] \leq \frac{0.1}{k_f} \cdot \frac{k_{r(\gamma^n)}k_{r(\beta^m)} + k_{r(\gamma^n)}k_b + k_{r(\beta^m)}k_b}{k_b + k_{r(\beta^m)}}$$

For situations where $m > 6$, $k_{r(\beta^m)} < k_b$, and the critical concentration can be approximated as $\frac{0.1\cdot(k_{r(\gamma^n)}+k_{r(\beta^m)})}{k_f}$. The numerator can be approximated as $0.1 \cdot \max(k_{r(\gamma^n)}, k_{r(\beta^m)}) = 6 \cdot 10^{5-\min(m,n)}$ s^{-1}, and the critical concentration is approximated as: $2 \cdot 10^{-1-\min(m,n)}$ M.

For situations where $m \leq 6$, $k_{r(\beta^m)} > k_b$, and the critical concentration can be approximated as $\frac{0.1\cdot(k_{r(\gamma^n)}+k_b)}{k_f}$. When $n > 6$, the critical concentration is approximated as $\frac{0.1k_b}{k_f} \approx 3 \cdot 10^{-8}$ M. When $n \leq 6$, the critical concentration is approximated as $\frac{0.1k_{r(\gamma^n)}}{k_f} = 2 \cdot 10^{-1-n}$ M.

S6. Matlab script for generating BM rate constant based on toehold energies

The follows shows a script for calculating the BM rate constant and the c_{crit} values for a toehold exchange reaction, taking the two toehold energies, the branch migration length, the temperature, and the energy model as inputs.

```
function output = BM_rate(input)

%input format:  [(invading toehold energy), (incumbent toehold energy),
%               (branch migration length), (temperature),  (energy model)]
%
%output format: [(BM rate constant), (critical concentration)]
%
%Assumes toeholds composition has all 4 bases (i.e. kf = 3.5e6)
%
%Energy model: 1 = default, 2 = Pyshnyi, 3 = Nupack + Protozanova,
%              4 = Nupack(dangles = all), 5 = Owczarzy
```

```
%
% NOTE:  User must manually add 2.38 kcal/mol for the input energies to NUPACK's
%        mole fraction energies

invading_energy = input(1);
incumbent_energy = input(2);
BM_length = input(3);
temperature = input(4)+273.15;
energy_model = input(5);

kf = 3.5e6;

if (energy_model == 1)
    kb = 1 * 400 / (BM_length * BM_length);
end
if (energy_model == 2)
    kb = 0.5 * 400 / (BM_length * BM_length);
end
if (energy_model == 3)
    kb = 1 * 400 / (BM_length * BM_length);
end
if (energy_model == 4)
    kb = 7 * 400 / (BM_length * BM_length);
end
if (energy_model == 5)
    kb = 4 * 400 / (BM_length * BM_length);
end

invading_offrate = kf * exp(invading_energy * 4180 / temperature / 8.314) * (2 / BM_length);
incumbent_offrate = kf * exp(incumbent_energy * 4180 / temperature / 8.314) * (2 / BM_length);

%BM rate constant
BM_rate = kf * kb * incumbent_offrate / (invading_offrate * incumbent_offrate ...
    + kb * invading_offrate + kb * incumbent_offrate)

%critical concentration
c_crit = (0.1 / kf) * (invading_offrate * incumbent_offrate + kb * invading_offrate ...
    + kb * incumbent_offrate) / (incumbent_offrate + kb)

output = [BM_rate, c_crit];
```

3-S7. Addendum: Derivation of BM rate constants with reverse reaction.

In the derivation of the BM rate constant in section 3-S1, we ignored the reverse reaction Y + L(m,n) $\rightarrow$ X(m,n) + S for simplicity of derivation and because it reflected experimental conditions. Here, we show that the BM rate constants derived in the analysis of the full reaction model including that reverse reaction is actually identical to that derived earlier. Recall that the net toehold exchange reaction can be expressed as:

$$\mathrm{X}(m,n) + \mathrm{S} \underset{k_{(\gamma^n,\beta_m,\beta^m)}}{\overset{k_{(\beta^m,\beta_m,\gamma^n)}}{\rightleftharpoons}} \mathrm{Y} + \mathrm{L}(m,n)$$

The expressions for the forward and reverse BM rate constants $k_{(\beta^m,\beta_m,\gamma^n)}$ and $k_{(\gamma^n,\beta_m,\beta^m)}$ must capture the kinetics of the overall reaction:

$$\frac{d[\mathrm{Y}]}{dt} = \frac{d[\mathrm{L}(m,n)]}{dt} = k_{(\beta^m,\beta_m,\gamma^n)}[\mathrm{X}(m,n)][\mathrm{S}] - k_{(\gamma^n,\beta_m,\beta^m)}[\mathrm{Y}][\mathrm{L}(m,n)] \quad (20)$$

$$\frac{d[\mathrm{X}(m,n)]}{dt} = \frac{d[\mathrm{S}]}{dt} = k_{(\gamma^n,\beta_m,\beta^m)}[\mathrm{Y}][\mathrm{L}(m,n)] - k_{(\beta^m,\beta_m,\gamma^n)}[\mathrm{X}(m,n)][\mathrm{S}] \quad (21)$$

From the three-step model, the rates of production of Y and S are expressed as:

$$\frac{d[\mathrm{Y}]}{dt} = k_{r(\beta^m)}[J] - k_f[\mathrm{Y}][\mathrm{L}(m,n)] \quad (22)$$

$$\frac{d[\mathrm{S}]}{dt} = k_{r(\beta^n)}[I] - k_f[\mathrm{X}(m,n)][\mathrm{S}] \quad (23)$$

We use QSS to analyze the production rate of Y and S from the three-step model. Under QSS, the steady state conditions on intermediates I and J are:

$$\frac{d[\mathrm{I}]}{dt} = k_f[\mathrm{X}(m,n)][\mathrm{S}] + k_b[\mathrm{J}] - k_b[\mathrm{I}] - k_{r(\gamma^n)}[\mathrm{I}] \approx 0 \quad (24)$$

$$\frac{d[\mathrm{J}]}{dt} = k_b[\mathrm{I}] + k_f[\mathrm{Y}][\mathrm{L}(m,n)] - k_b[\mathrm{J}] - k_{r(\beta^m)}[\mathrm{J}] \approx 0 \quad (25)$$

Rearranging eq. (25),

$$[\mathrm{I}] = \frac{k_b[\mathrm{J}] + k_{r(\beta^m)}[\mathrm{J}] - k_f[\mathrm{Y}][\mathrm{L}(m,n)]}{k_b}$$

Substituting this expression for [I] back into equation (24),

$$\begin{aligned}
0 &= k_f[\mathrm{X}(m,n)][\mathrm{S}] + k_b[\mathrm{J}] - \Big(k_b[\mathrm{J}] + k_{r(\beta^m)}[\mathrm{J}] - k_f[\mathrm{Y}][\mathrm{L}(m,n)]\Big)\frac{k_b + k_{r(\gamma^n)}}{k_b} \\
0 &= k_f[\mathrm{X}(m,n)][\mathrm{S}] + \Big(k_f + \frac{k_f k_{r(\gamma^n)}}{k_b}\Big)[\mathrm{Y}][\mathrm{L}(m,n)] - \Big(k_{r(\beta^m)} + k_{r(\gamma^n)} + \frac{k_{r(\beta^m)}k_{r(\gamma^n)}}{k_b}\Big)[\mathrm{J}] \\
[\mathrm{J}] &= \frac{k_f k_b[\mathrm{X}(m,n)][\mathrm{S}] + (k_f k_b + k_f k_{r(\gamma^n)})[\mathrm{Y}][\mathrm{L}(m,n)]}{k_b k_{r(\beta^m)} + k_b k_{r(\gamma^n)} + k_{r(\gamma^n)}k_{r(\beta^m)}} \quad .
\end{aligned}$$

Substituting this expression for [J] into equation (22),

$$\begin{aligned}
\frac{d[\mathrm{Y}]}{dt} &= k_{r(\beta^m)}[\mathrm{J}] - k_f[\mathrm{Y}][\mathrm{L}(m,n)] \\
&= \frac{k_f k_b k_{r(\beta^m)}[\mathrm{X}(m,n)][\mathrm{S}] + (k_f k_b k_{r(\beta^m)} + k_f k_{r(\gamma^n)}k_{r(\beta^m)})[\mathrm{Y}][\mathrm{L}(m,n)]}{k_b k_{r(\beta^m)} + k_b k_{r(\gamma^n)} + k_{r(\gamma^n)}k_{r(\beta^m)}} - k_f[\mathrm{Y}][\mathrm{L}(m,n)] \\
&= \frac{k_{r(\beta^m)}k_f k_b}{k_{r(\gamma^n)}k_{r(\beta^m)} + k_{r(\gamma^n)}k_b + k_{r(\beta^m)}k_b}[\mathrm{X}(m,n)][\mathrm{S}] - \frac{k_{r(\gamma^n)}k_f k_b}{k_{r(\gamma^n)}k_{r(\beta^m)} + k_{r(\gamma^n)}k_b + k_{r(\beta^m)}k_b}[\mathrm{Y}][\mathrm{L}(m,n)] \quad .
\end{aligned}$$

Combining this result with equation (20), we derive the BM rate constants $k_{(\beta^m,\beta_m,\gamma^n)}$ and $k_{(\gamma^n,\beta_m,\beta^m)}$:

$$k_{(\beta^m,\beta_m,\gamma^n)} = \frac{k_{r(\beta^m)}k_f k_b}{k_{r(\gamma^n)}k_{r(\beta^m)} + k_{r(\gamma^n)}k_b + k_{r(\beta^m)}k_b}$$

$$k_{(\gamma^n,\beta_m,\beta^m)} = \frac{k_{r(\gamma^n)}k_f k_b}{k_{r(\gamma^n)}k_{r(\beta^m)} + k_{r(\gamma^n)}k_b + k_{r(\beta^m)}k_b} \quad .$$

Solving for steady state [I] and substituting into equation (23) yields the same rate constant expressions.

[1] Kallenbach, N. R.; Ma, R. I.; Seeman, N. C. *Nature* **1983,** *305*, 829.

[2] Rothemund, P. *Nature* **2006,** *440*, 297.

[3] Winfree, E.; Liu, F.; Wenzler, L. A.; Seeman, N. C. *Nature* **1998,** *394*, 539.

[4] He, Y.; Ye, T.; Su, M.; Zhang, C.; Ribbe, A. E.; Jiang, W.; Mao, C. *Nature* **2008,** *452*, 198.

[5] Yurke, B.; Turberfield, A. J.; Mills, A. P.; Simmel, F. C.; Neumann, J. L. *Nature* **2000,** *406*, 605.

[6] Dirks, R. M.; Pierce, N. A. *Proc. Nat. Acad. Sci.* **2004,** *101*, 15275.

[7] Pei, R.; Taylor, S. K.; Stefanovic, D.; Rudchenko, S.; Mitchell, T. E.; Stojanovic, M. N. *J. Am. Chem. Soc.* **2006,** *128*, 12693.

[8] Green, S. J.; Bath, J.; Turberfield, A. J. *Phys. Rev. Lett.* **2008,** *101*, 238101.

[9] Yin, P.; Choi, H. M. T.; Calvert, C. R.; Pierce, N. A. *Nature* **2008,** *451*, 318.

[10] Shin, J. S.; Pierce, N. A. *J. Am. Chem. Soc.* **2004,** *126*, 10834.

[11] Sherman, W. B.; Seeman, N. C. *Nano Letters* **2004,** *4*, 1203.

[12] Omabegho, T.; Sha, R.; Seeman, N. C. *Science* **2009,** *324*, 67.

[13] Stojanovic, M. N.; Semova, S.; Kolpashchikov, D.; Macdonald, J.; Morgan, C.; Stefanovic, D. *J. Am. Chem. Soc.* **2005,** *127*, 6914.

[14] Penchovsky, R.; Breaker, R. R. *Nat. Biotech.* **2005,** *23*, 1424.

[15] Seelig, G.; Soloveichik, D.; Zhang, D. Y.; Winfree, E. *Science* **2006,** *314*, 1585.

[16] Zhang, D. Y.; Turberfield, A. J.; Yurke, B.; Winfree, E. *Science* **2007,** *318*, 1121.

[17] Dirks, R. M.; Bois, J. S.; Schaeffer, J. M.; Winfree, E.; Pierce, N. A. *SIAM Rev.* **2007,** *49*, 65.

[18] Watson, J. D.; Crick, F. H.; *Nature* **1953,** *171*, 737.

[19] Zuker, M. *Nucleic Acids Res.* **2003,** *31*, 3406.

[20] SantaLucia, J. *Proc. Natl. Acad. Sci.* **1998,** *95*, 1460.

[21] SantaLucia, J.; Hicks, D. *Annu. Rev. Biophys. Biomol. Struct.* **2004,** *33*, 415.

[22] Bommarito, S.; Peyret, N.; SantaLucia, J. *Nuc. Acid Res.* **2000,** *28*, 1929.

[23] Peyret, N. *Prediction of Nucleic Acid Hybridization: Parameters and Algorithms* **2000,** Doctoral thesis, Wayne State University.

[24] Tan, Z. J.; Chen, S. J. *Biophys. J.* **2006,** *90*, 1175.

[25] Owczarzy, R.; Moreira, B. G.; You, Y.; Behlke, M. A.; Walder, J. A. *Biochemistry* **2008,** *47*, 5336.

[26] Protozanova, E.; Yakovchuk, P.; Frank-Kamenetskii, M. D. *J. Mol. Biol.* **2004,** *342*, 775.

[27] Pyshnyi, D. V.; Ivanova, E. M. *Russian Chemical Bulletin* **2002,** *51*, 1145.

[28] Vasiliskov, V. A.; Prokopenko, D. V.; Mirzabekov, A. D. *Nuc. Acid Res.* **2001,** *29*, 2303.

[29] Pyshnyi, D. V.; Ivanova, E. M. *Nucleosides, Nucleotides, and Nucleic Acids* **2004,** *23*, 1057.

[30] Petruska, J.; Goodman, M. F. *J. Biological Chemistry* **1995,** *270*, 746.

[31] Cooper, A.; Johnson, C. M.; Lakey, J. H.; Nollmann, M. *Biophysical Chemistry* **2001,** *93*, 215.

[32] Bloomfield, V. A.; Crothers, D. M.; Tinoco, I. Jr. *Nucleic Acids: Structures, Properties, and Functions* **2000,** University Science Books, Sausalito, CA.

[33] Zheng, M.; Jagota, A.; Semke, E. D.; Diner, B. A.; Mclean, R. S.; Lustig, S. R.; Richardson, R. E.; Tassi, N. G. *Nature Materials* **2003,** *2*, 338.

[34] Maune, H. T.; Han S.; Barish R. D.; Rothemund, P. W. K.; Goddard, W. A.; Bockrath, M.; Winfree, E. "Self-assembly of Carbon Nanotubes into Two-dimensional Geometries using DNA Origami Templates." *Nature Nanotech.* **2009,** (accepted).

[35] Elghanian, R.; Storhoff, J. J.; Mucic, R. C.; Letsinger, R. L.; Mirkin, C. A. *Science* **1997,** *277*, 1078.

[36] Park, S. Y.; Lytton-Jean, A. K. R.; Lee, B.; Weigand, S.; Schatz, G. C.; Mirkin, C. A. *Nature* **2008,** *451*, 553.

[37] Aldaye, F. A.; Palmer, A. L.; Sleiman, H. F. *Science* **2008,** *321*, 1795.

[38] Gianneschi, N. C.; Ghadiri, M. R. *Angew. Chem. Int. Ed.* **2007,** *46,* 3955.

[39] Gartner, Z. J.; Kana, M. W.; Liu, D. R. *J. Am. Chem. Soc.* **2002,** *124*, 10304.

[40] Zhu L.; Lukeman, P. S.; Canary, J. W.; Seeman, N. C. *J. Am. Chem. Soc.* **2003,** *125*, 10178.

[41] Zhang, D. Y.; Winfree, E. *J. Am. Chem. Soc.* **2008,** *130*, 13921.

[42] Panyutin, I. G.; Hsieh, P. *J. Mol. Biol.* **1993,** *230*, 413.

[43] Radding, C. M.; Beattie, K. L.; Holloman, W. K.; Wiegand, R. C. *J. Mol. Biol.* **1977,** *116*, 825.

[44] Panyutin, I. G.; Hsieh, P. *Proc. Nat. Acad. Sci.* **1994,** *91*, 2021.

[45] Morrison, L. E.; Stols, L. M. *Biochemistry* **1993,** *32*, 3095.

[46] Green, C.; Tibbetts, C. *Nuc. Acid Res.* **1981,** *9*, 1905.

[47] Weinstock, P.; Wetmur, J. *Nuc. Acid Res.* **1990,** *18*, 4207.

[48] Hagiya, M.; Yaegashi, S.; Takahashi, K. *Nanotechnology: Science and Computation*, **2006,** 293.

[49] Frezza, B. M.; Cockroft, S. L.; Ghadiri, M. R. *J. Am. Chem. Soc*, **2007,** *129*, 14875.

[50] Bois, J. S.; Venkataraman, S.; Choi, H. M. T; Spakowitz, A. J.; Wang, Z. G.; Pierce, N. A. *Nuc. Acid Res.* **2005,** *33*, 4090.

[51] Seelig, G.; Yurke, B.; Winfree, E. *J. Am. Chem. Soc.* **2006,** *128*, 12211.

[52] Turberfield, A. J.; Mitchell, J. C.; Yurke, B.; Mills, A. P.; Blakey, M. I.; Simmel, F. C. *Phys. Rev. Lett.* **2003,** *90*, 118102.

[53] Yurke, B.; Mills, A. P. *Genet. Prog. Evol. Mach.* **2003,** *4*, 111.

[54] Gao, Y.; Wolf, L. K.; Georgiadis, R. M. *Nucleic Acids Research* **2006,** *34*, 3370.

[55] Green, S. J.; Lubrich, D.; Turberfield, A. J. *Biophys. J.* **2006,** *91*, 2966.

[56] Soloveichik, D.; Seelig, G.; Winfree, E. *Proceedings of the 14th Annual Conference on DNA computing*, **2008,** Prague, Czech Republic.

[57] Qian, L.; Winfree, E. *Proceedings of the 14th Annual Conference on DNA computing*, **2008,** Prague, Czech Republic.

[58] Cardelli, L. *Proceedings of the 15th Annual Conference on DNA computing*, **2009,** Fayetteville, Arkansas.

[59] Rinker, S.; Ke, Y.; Liu, Y.; Chhabra, R.; Yan, H. *Nat. Nanotech.* **2008,** *3*, 418.

[60] Cantor, C. R.; Schimmel, P. R. *Biophysical Chemistry: Part III: The Behavior of Biological Macromolecules* **2000,** W. H. Freeman, San Francisco, CA.

[61] Dimitrov, R. A.; Zuker M. *Biophysical Journal* **2004,** *87*, 215.

[62] Segel, L. A.; Slemrod, M. *SIAM Rev.* **1989,** *31* 446.

[63] Dirks, R. M.; Lin, M.; Winfree, E.; Pierce, N. A. *Nucleic Acids Res.* **2004,** *32*, 1392.

[64] Reynaldo, L. P.; Vologodskii, A. V.; Neri, B. P.; Lyamichev, V. I. *J. Mol. Bio.* **2000,** *297*, 511.

[65] Puglisi, J. D.; Tinoco, I. *Methods in Enzymology* **1989,** *180*, 304.

[66] Marras, S. A.; Kramer, F. R.; Tyagi, S. *Nucleic Acids Research* **2002,** *30*, 122.

[67] Flamm, C.; Hofacker, I. L. *Monatsh. Chem.* **2008,** *139*, 447.

[68] Sun, W.; Mao, C.; Liu, F.; Seeman, N. C. *J. Mol. Biol.* **1998,** *282*, 59.

[69] Yuan, B.; Zhuang, X.; Hao, Y.; Tan, Z. *Chem. Comm.* **2008,** *48*, 6600.

[70] Feller, W. *An Introduction to Probability Theory and Its Applications* **1958,** John Wiley and Sons, Singapore, Singapore, chapter 5.

[71] Based on the dependence of k_{f1} on the sequence composition of the toehold, we believe that k_{f2} is likely closest in value to $k_{f1} = 3.5 \cdot 10^6$ M^{-1} s^{-1} for inputs $X(m,n)$ with toeholds containing both strong G-C pairs and weak A-T pairs. However, we have no way of knowing the actual value of k_{f2} without performing a new series of experiments on the β^m toehold, similar to those shown in Figure 3-3.

[72] To take a simple example, consider when the invading and incumbent toeholds are both 4 nt, but the branch migration domain is 1000 nt long. Molecules will spend a significant amount of time in a three-stranded complex, with the branch migration junction vacillating near the middle of the branch migration domain. Our model, with the $\frac{2}{b-m}$ correction, will account for the high multiplicity of branch migration microstates.

[73] ODEs used for fitting $k_{\{m,n\}}$ simulated the two reactions $X(m,n) + S \rightarrow Y + L(m,n)$ and $Y + R \rightarrow F$. The reverse reaction $Y + L(m,n) \rightarrow S + X(m,n)$ is not included because (1) it is desirable to fit one parameter at a time and (2) because the value of the reverse rate constant $\kappa_{\{n,m\}}$ is constrained poorly by data from experiments of the type shown in Fig. 3-3A. Furthermore, the concentration of the reporter R is in 3x excess over that of the substrate S and the rate constant of the reaction between Y and R is high; consequently, it is expected that the reaction between Y and R will dominate that between Y and $L(m,n)$. Fitting $k_{\{m,n\}}$ to a full model including the reverse reaction with its rate constant $\kappa_{\{m,n\}}$ set to the value predicted by our model yielded values very similar to those presented in the paper (difference of no more than 10%).

Chapter 4: Amplification and Transduction of DNA Signals

One crucial element in both biology and engineering is the idea of amplification. In biology, one copy of DNA can produce thousands of copies of mRNA, each of which can in turn produce thousands of copies of protein. In electrical engineering, the key invention of the twentieth century was the transistor, which amplifies a small base current into a large emitter current (in the case of an NPN bipolar transistor). In order for nucleic acid engineering to become more than just a toy science, a similar amplification method needs to be developed–something like a biomolecular transistor.

In chemistry, catalysis reactions form gain of a sort; each catalyst molecule turns over multiple copies of product. The problem with using chemical catalysis as the basis for molecular engineering is that chemical catalysis is extremely difficult to design. Arbitrary chemical molecules cannot in general be made to catalyze the production of other arbitrary chemical molecules. But in working with DNA, we *can* engineer arbitrary catalysis reactions... as long as we slightly modify the definition of catalysis.

The work below demonstrates a reliable method for constructing DNA catalysis reactions, in which a target DNA molecule of arbitrary sequence can be made to catalytically release a product DNA molecule of independent sequence from a frustrated multi-stranded complex. This amplification mechanism uses the toehold exchange mechanism described in

Chapter 2, and is unconventional for chemical catalysis in that it does not make or break any covalent bonds. However, as we show below, these amplification components can be integrated in a modular fashion into larger DNA circuits... circuits that one day could control *in vivo* behavior of cells and organisms. As proof-of-concept experiments, we demonstrated a serial cascade of two amplification elements that quadratic signal gain, and is reminiscent of signal transduction cascades (such as that of MAP kinases). We also demonstrate a feedback network exhibiting exponential concentration growth; this was particularly exciting because it was the first work to show a generalized construction of autocatalysts using purely rationally designed components.

I came up with the initial idea for this work while in Andrew Turberfield's laboratory at the University of Oxford in 2005. I then did some preliminary experiments, which did not behave particularly well due to uncharacterized fluorophore-quencher interactions. I continued experimentally working on this project upon starting graduate school in Erik Winfree's lab at Caltech, where it was finally finished in Spring 2007. Erik Winfree, Bernard Yurke, and Andrew Turberfield helped model the system, analyze the data, and revise the manuscript. Parts of this work were presented at the 12th International DNA Computing conference in Seoul, South Korea. This work was published in full as:

Zhang, David Yu; Turberfield, Andrew J.; Yurke, Bernard; Winfree, Erik. "Engineering Entropy-Driven Reactions and Networks Catalyzed by DNA." *Science* 318: 1121-1125 (2007).

Artificial biochemical circuits are likely to play as large a role in biological engineering as electrical circuits have played in the engineering of electromechanical devices. Toward that end, nucleic acids provide a designable substrate for regulation of biochemical reactions. However, it has been difficult to incorporate signal amplification components. We introduce a design strategy that allows a specified input oligonucleotide to catalyze the release of a specified output oligonucleotide, which in turn can serve as a catalyst for other reactions.

This reaction, which is driven forward by the configurational entropy of the released molecule, provides an amplifying circuit element that is simple, fast, modular, composable, and robust. We have constructed and characterized several circuits that amplify nucleic acid signals, including a feedforward cascade with quadratic kinetics and a positive feedback circuit with exponential growth kinetics.

The development of modular biochemical circuit elements poses several challenges. First, distinct signals must be carried by distinct chemical species, motivating the use of information-carrying molecules whose sequences can be used to encode signal identity. Second, "wiring up" a gate to specified inputs and outputs involves the design and synthesis of new molecules; this calls for modular gate designs. Third, to create gates with signal gain, a fast and robust catalytic mechanism must be identified and coupled to a suitable energy source. Fourth, it must be possible to construct circuits of arbitrary complexity that can produce an unlimited variety of dynamical behaviors. Finally, there should be no leak or crosstalk between distinct signals and gates. It is difficult to meet all these challenges simultaneously.

Nucleic acids are attractive for this purpose because the combinatorial sequence space allows for an enormous diversity of signal carriers, while the predictability and specificity of Watson-Crick base pairing facilitates the design of gate architectures. The RNA world hypothesis further suggests that sophisticated biochemical organization can be achieved with nucleic acids alone [1], and indeed nucleic acids have been shown to be a versatile construction material for engineering molecular structures and devices [2, 3], including catalytic [4–8] and logical [9–12] control elements and circuits [13–17]. Engineering (deoxy)ribozyme-based logic gates has been remarkably effective, resulting in systems containing over 100 gates operating independently in parallel [10] as well as systems demonstrating cascading of a signal between two gates [13, 15, 16]. Hybridization-based systems, usually driven by the energy of base-pair formation, have proven especially suitable for cascading signals, as demonstrated by a circuit five layers deep [17]. That work, relying primarily on non-catalytic logic gates, identified amplification and signal gain as essential for scaling up to large cascaded circuits.

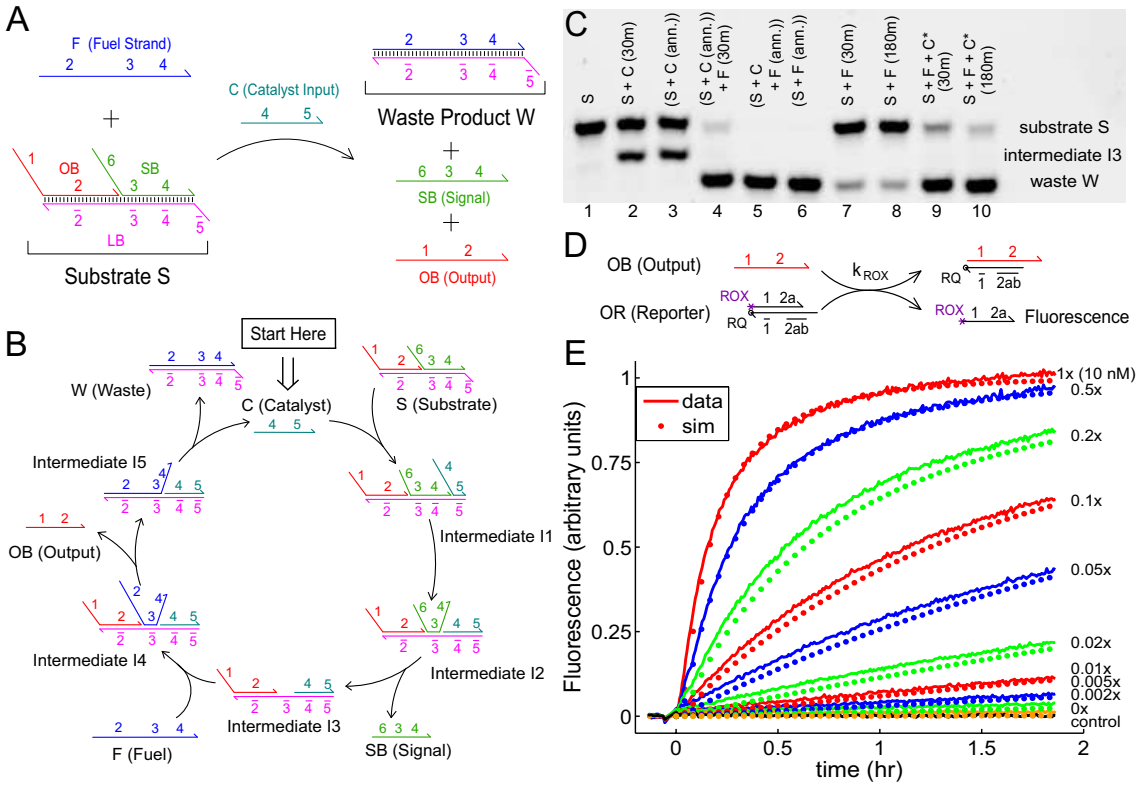


FIG. 4-1: The entropy-driven reaction. **(A)** System components. Number labels denote functional domains, continuous stretches of DNA that act as units in binding. Domain $\bar{x}$ is the complement of (and will hybridize to) domain x. **(B)** The proposed catalytic pathway. Reverse reactions are also present and modeled, with the exception of $I5 + OB \rightarrow I4$ which occurs at a negligible rate. **(C)** Analysis by PAGE (12% native gel) of the reaction mechanism. Unless otherwise noted, all experiments were performed at 25 °C in Tris-acetate (TE) buffer supplemented with 12.5 mM $MgCl_2$. Here, $[S] = [F] = 200$ nM. $[C] = 200$ nM, except where $C*$ denotes 20 nM. "(ann.)" denotes that species were annealed. "(30m)" denotes that the reaction occurred for 30 minutes. See Fig. 4-S5 for full gel including control lanes [22]. **(D)** Fluorescent reporter strategy. ROX denotes the carboxy-X-rhodamine fluorophore, and RQ denotes the Iowa Black Red Quencher. Domain 2 is subdivided into $2a$, $2b$, and $2c$; $2ab$ consists of $2a$ and $2b$ (see Table 4-1). **(E)** Demonstration of catalysis. Different amounts of C were introduced into the system at $t \approx 0$. Here, $[S] =$ 10 nM = 1x, $[F] = 13$ nM, $[OR] = 30$ nM. Fluorescence (in all figures) is reported in units such that 0.0 is the background fluorescence of the quenched reporter, and 1.0 is the fluorescence of approximately 10 nM of triggered reporter. The control trace (black) shows the reaction with no substrate S and no catalyst C . Dotted lines show curves calculated using the reduced reaction model.

Here, we provide a solution to this problem.

The entropy-driven catalytic gate presented here is significantly simpler than previous hybridization-based designs, yet it is faster, better understood, and more modular. The net reaction is shown in Fig. 4-1A: fuel strand (F) reacts with the three-stranded substrate complex (S), displacing output and signal strands (OB and SB) from linker strand (LB) to form waste complex (W). The total number of base pairs in the reactants and products

is unchanged; the reaction is driven forward thermodynamically by the entropic gain of the liberated molecules. Fuel, signal, catalyst, and output are all single-stranded DNA molecules that can be of similar lengths, thus each molecule may serve multiple roles within a network. For example, the output of one gate may serve as the input to another. Importantly, catalyst C and output OB may be entirely independent in sequence [18]; this modularity implies that a catalytic gate can be designed to act at any point within a pre-existing circuit. Unlike previous hybridization-based catalyst systems, the reaction design does not require unusual secondary structures such as pseudoknots and kissing loops. Undesired interactions can be avoided by design [19–21], resulting in reliable and predictable circuit behavior.

Strands are conceptually subdivided into functional domains (number labels in Fig. 4-1) whose sequences determine the pattern of interactions between circuit components. (Domain sequences are given in Table 4-1; see text 4-S1 for design details [22].) The domains can be grouped by purpose: 3 and 5 are termed toehold domains, while 1, 2, 4, and 6 are termed specificity domains. Toehold domains are short enough to bind only fleetingly in the absence of additional binding (and need not be unique), but they greatly accelerate the initiation of strand displacement reactions [23]. Specificity domains ensure specific interactions (even a single mismatch can slow down branch migration substantially [24]) and uniquely determine the identities of the catalyst and output molecules. The lengths of the toehold domains determine kinetics and need to be between roughly 4 and 10 nucleotides (nt), but the specificity domains may be of any length sufficient to ensure thermal stability. Domains 1 and 6 of OB and SB, respectively, are inert while their respective toeholds are sequestered in S.

Catalytic activity has two main characteristics: the speedup of the target reaction and the re-release of the catalyst to allow for multiple turnover. To achieve this, we utilize toehold exchange (see Fig. 4-1B): C first binds to the single-stranded toehold domain $\bar{5}$ on S to form four-stranded intermediate $I1$, which then rearranges (by branch migration) to form $I2$. The binding between toehold domains 3 and $\bar{3}$ is too weak to keep SB attached, so $I2$ spontaneously dissociates into SB and $I3$. Newly-exposed $\bar{3}$ then facilitates the binding of fuel F, resulting in $I4$, which then quickly rearranges to release OB and $I5$. Finally,

Dom.	Sequence	Length (nt)
1	5'- CTTTCCTACA -3'	10
2a	5'- CCTACG -3'	6
2b	5'- TCTCCA -3'	6
2c	5'- ACTAACTTACGG -3'	12
3	5'- CCCT -3'	4
4	5'- CATTCAATACCCTACG -3'	16
5	5'- TCTCCA -3'	6
6	5'- CCACATACATCATATT -3'	16

Table 4-1: Domain sequences of basic catalytic reaction

$I5$ rearranges so that C is attached only by the binding of 5 and $\bar{5}$, which spontaneously dissociates to leave waste W and regenerate catalyst C. Note that the reaction mechanism presented here, based on branch migration and driven by entropy, differs from the traditional view of catalysis in biological organisms in that it requires no enzymes and alters no covalent bonds.

It is important to ensure that alternative interactions do not interfere with intended gate functions. To this end, a key design principle is that complements of the specificity domains never appear in their single-stranded form. Except at toeholds, no two molecules interact with each other via complementary single-stranded domains. The catalytic gate is therefore expected to function for most choices of domain sequences lacking significant secondary structure and spurious mutual interactions [19–21].

In Fig. 4-1C, polyacrylamide gel electrophoresis (PAGE) is used to verify the catalytic pathway (see Text 4-S2 for materials and methods [22]). By reacting substrate S (purified by gel) and catalyst C in the absence of fuel F, we prevent the reaction from progressing past intermediate $I3$. The amount of $I3$ produced after 30 minutes (lane 2) is almost identical to that at equilibrium, as assessed by annealing the reaction components (lane 3). This suggests that all reactions up to $I3$ are fast on this time scale. Similarly, the subsequent reaction between $I3$ and F is also fast (lanes 4-5). The complete system behaves as expected: the uncatalyzed reaction is slow (lanes 7-8), and a sub-stoichiometry quantity (0.1x) of C enables the reaction to proceed rapidly to near-completion (lanes 9-10).

In order to measure the time course of the catalyzed reaction by means of a fluorescent reporter without interference from fluorophore/quencher interactions (see [25] and text 4-S3 [22]), we utilize an indirect reporter complex OR. OR reacts stoichiometrically with

output OB to separate a fluorophore-labeled strand from a quencher-labeled strand, thereby increasing fluorescence (Fig. 4-1D). The rate constant for the reporter system was measured to be $k_{ROX} = 4 \cdot 10^5 M^{-1}s^{-1}$ (see text 4-S4 [22]). Since $[OR]$= 30 nM while $[S]$= 10 nM, the reporter complex remains significantly in excess and the reporting delay should remain less than 100 s – short compared to the time of the catalyzed reactions. OR does not react significantly with S, since there are no single-stranded toeholds to initiate interaction. Measurements of the kinetics of the catalyzed reaction over a 500-fold range of catalyst concentration are shown in Fig. 4-1E.

We modeled this system using the reduced reaction set shown below.

$$
\begin{array}{ccccc}
S + F & \stackrel{k_0}{\rightarrow} & SB + OB + W & & k_0 = 2.3 \cdot 10^1 M^{-1}s^{-1} \\
 & & & & \\
S + C & \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} & I3 + SB & & k_1 = 6.5 \cdot 10^5 M^{-1}s^{-1} \\
I3 + F & \stackrel{k_2}{\rightarrow} & I5 + OB & & k_2 = 4.2 \cdot 10^5 M^{-1}s^{-1} \\
I5 & \underset{k_{-3}}{\overset{k_3}{\rightleftharpoons}} & C + W & & k_3 = 4 \cdot 10^{-3} s^{-1} \text{ (fitted)} \\
OB + OR & \stackrel{k_{ROX}}{\rightarrow} & \text{Fluorescence} & & k_{ROX} = 4 \cdot 10^5 M^{-1}s^{-1}
\end{array}
$$

The first reaction shown models the uncatalyzed (leak) reaction. Intermediate steps in branch-migration reactions are omitted because they are relatively fast at experimental concentrations (see text 4-S5 [22]) [26], and because intermediates $I1$, $I2$, and $I4$ are not observed in PAGE analysis of reactants and products (Fig. 4-1B). Noting the approximate symmetry between the corresponding reactions, we assume that $k_{-3} = k_1$ and $k_{-1} = k_2$. The rate constants k_0, k_1, and k_2 were measured individually (see Fig. 4-S4 [22]); k_3 was fit to the data of Fig. 4-1E. The time course of the catalyzed reaction over a wide range of catalyst concentrations is accurately reproduced by this reduced system of rate equations (Fig. 4-1E). According to this model, the addition of catalyst can accelerate the reaction by over four orders of magnitude ($k_2/k_0 = 1.8 \cdot 10^4$).

In the net reaction, each base pair that is broken is replaced by another of the same type, so the net free energy change from base-pairing interactions should be insignificant. The reaction is driven by the gain in configurational entropy corresponding to the liberation of OB and SB at the cost of localizing F. To confirm the dominance of this entropic driving

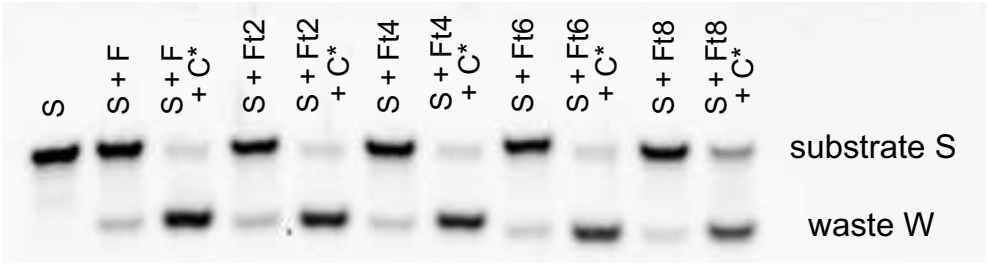


FIG. 4-2: Verification of entropic driving force. Analysis by PAGE (12% native gel) of reactions using truncated fuel strands. $[S] = [F] = 200$ nM. $[C] = 20$ nM, as denoted by the * label. All reactions were run at 25 °C for 3 hours. "$Ft2$" denotes that 2 bases were truncated from the 5' end of fuel strand F.

force, we truncate F by removing up to 8 nt from its 5' end, making the products more and more thermodynamically disfavored. Nonetheless, in all cases the waste product is favored at equilibrium (Fig. 4-2; see text 4-S6 for discussion on entropy and free energies [22]). The thermodynamic driving force, being dominated by center-of-mass configurational entropy of released molecules, is somewhat robust to environmental conditions such as temperature and salt concentrations that alter the strength of DNA hybridization (see Text 4-S7 and Fig. 4-S6 [22]).

To demonstrate cascaded circuit construction, a two-layer feedforward network was designed by introducing an upstream catalyst system whose output acts as the catalyst for the original system (Fig. 4-3A). For clarity, F, OB, etc. from Fig. 4-1 are relabeled $F1$, $OB1$, etc. Catalyst $C0$ catalyzes the production of $OB0$ (which contains a subsequence identical to C from Fig. 4-1), which in turn catalyzes the production of $OB1$. The concentration of upstream catalyst $C0$ is constant, so initially $[OB0]$ increases linearly with time, which causes $[OB1]$ to increase quadratically with time (Fig. 4-3B). Eventually, the substrates and fuels are depleted, and the reaction halts, giving rise to an overall sigmoidal shape to the fluorescence traces (Fig. 4-3B). The model previously used can be extended to predict the behavior of this feedforward circuit data (see text 4-S8 [22]).

This cascaded system can be used as an amplifier to detect small quantities of $C0$. Repeated fluorescence experiments show that we are able to distinguish reliably between 1 pM (0.0001x) catalyst $C0$ and 0x catalyst within 12 hours (Fig. 4-3B inset). This corresponds to a roughly 900-fold amplification of the input signal. (1 pM of catalyst triggered $\approx$ 900 pM

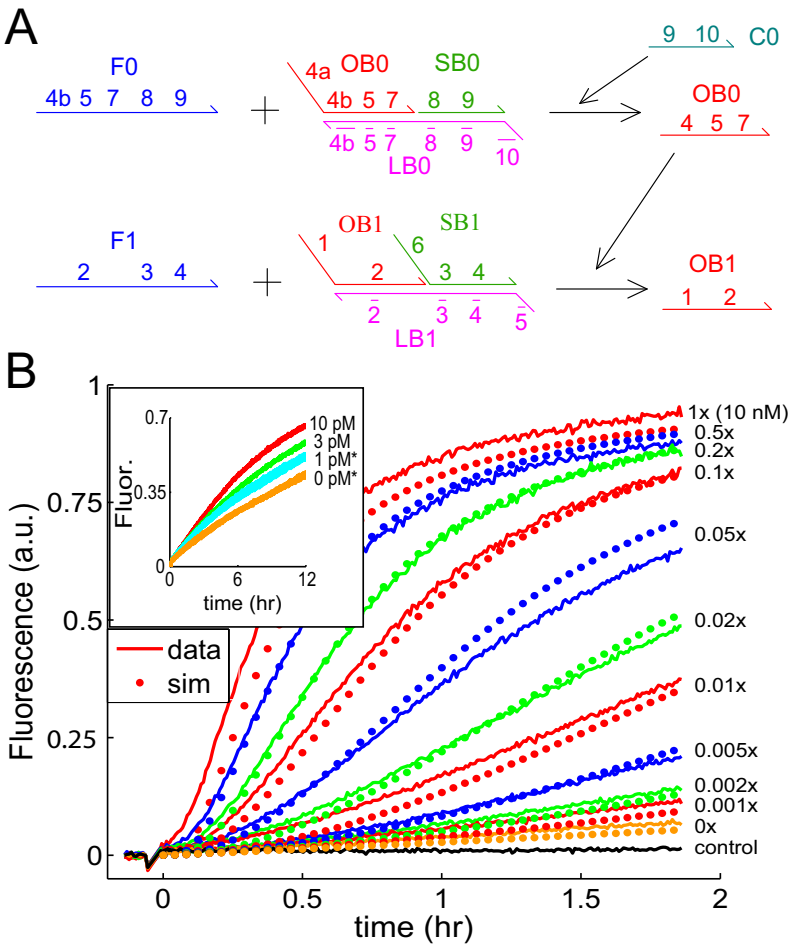


FIG. 4-3: A two-layer cascaded network. **(A)** Schematic. See Table 4-S2 for sequences of new domains [22]. **(B)** Kinetics. Indicated amounts of initial catalyst $C0$ were added at $t \approx 0$. Fluorescence derives from reporter complex OR (Fig. 4-1D) at 30 nM. Dotted lines show simulated traces; see text 4-S8 for details on reaction rates and modeling [22]. **(Inset)** Response to 0.0010x, 0.0003x, and 0.0001x catalyst. Asterisk (*) indicates that three independent reaction traces are shown. 1.0 fluorescence units corresponds to $\approx$ 10 nM of triggered reporter.

of reporter above the baseline set by the 0x reaction.) For comparison, 1 pM corresponds to approximately one molecule per eukaryotic cell volume. Repeated measurements of independent samples show less than 3% variability across all time points (see text 4-S9 [22]).

Feedback is another important feature of both biological regulatory networks and artificial control circuits. Exponential growth kinetics can be achieved by redesigning the reaction presented in Fig. 4-1 such that output OB contains catalyst C as a subsequence (Fig. 4-4A). The reaction is then autocatalytic. Fig. 4-4C shows the time course of this reaction for a wide range of catalyst concentrations. In a process dominated by initial ex-

ponential growth ($c \approx c_0 e^{\lambda t}$), the time to reach a threshold degree of completion depends logarithmically on the initial concentration c_0. Thus, a linear trend in a log-linear plot of initial concentration to time to half completion ($t_{1/2}$) is indicative of exponential growth. (Such plots are used as calibration standards for quantitation methods such as real-time PCR [27].) Fig. 4-4D shows that our autocatalytic system has this characteristic behavior, implying that exponential growth kinetics have in fact been achieved, and that the reaction is not significantly affected by product inhibition. Further confirmation comes from the quality of fit to the data of a model based on rate constants derived for the catalyst system of Fig. 4-1 (see text 4-S8 [22]).

We also demonstrate feedback in a two-layer circuit by redesigning $OB1$ so that it can in turn catalyze the $F0 + S0$ reaction (see text 4-S11 [22]). Feedback in this cross-catalytic system causes the concentrations of both $OB0$ and $OB1$ to grow exponentially at early times.

Largely because of their relevance to the origin of life and to the RNA world [1], autocatalytic and cross-catalytic self-replication reactions have been proposed and demonstrated previously [28]. However, such systems typically suffer from product inhibition and thus exhibit parabolic rather than exponential growth kinetics. Recent exceptions include cross-catalytic deoxyribozymogens [13] and catalyzed self-assembly [29] based on the hybridization chain reaction [14]; our autocatalyst system is faster than either. Reducing the spontaneous activity of the circuit, for example by improved purification of the substrate complex, is an important goal for increasing the sensitivity to the point that our autocatalyst could be used as an enzyme-free constant-temperature alternative to PCR for detecting known sequences.

For many applications in biotechnology, nucleic acid devices must remain functional in the presence of naturally occurring macromolecules. We therefore tested the autocatalyst system in the presence of an excess of mouse liver total RNA with rabbit reticulocyte lysate (Fig. 4-4E). Reactions proceeded to apparent completion with no more than a two-fold slowdown, and presence of a 3% trigger can still be detected.

The ability to construct larger circuits will enable the wide range of chemical circuit func-

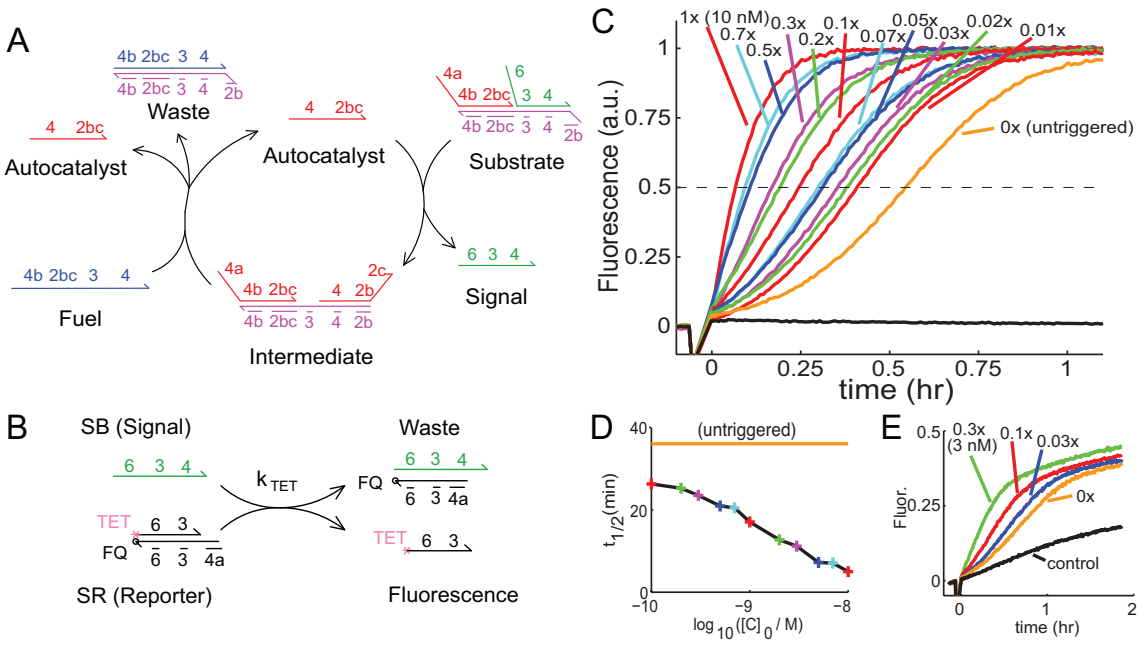


FIG. 4-4: The autocatalyst system. **(A)** Proposed reaction pathway. See Table 4-S2 for sequences of new domains [22]. **(B)** Reporter complex SR, used for monitoring autocatalytic and cross-catalytic reaction networks. TET denotes the tetrachlorofluorescein fluorophore. FQ denotes the Iowa Black Fluorescent Quencher. **(C)** Kinetics of autocatalysis. Indicated amounts of autocatalyst were added at $t \approx 0$. At 30 min, 1% (100 pM) was amplified 25-fold over the untriggered reaction. Reporter SR was present at 20 nM = 2x. Control sample (black) contained no fuel F. **(D)** Semi-log plot of time (in minutes) to half completion as a function of logarithm of trigger concentration (in molar). The orange line shows the time to half completion of the untriggered reaction. **(E)** Performance in the simultaneous presence of total RNA and unfractionated cell lysate. Total RNA present in solution was 10x (by mass) that of the sum of all relevant catalyst DNA. Active cell lysate, as would be used in an in vitro translation system, was added to be 1.1% by volume of total reaction (17 μL in 1500 μL). The control reaction did not contain any substrate S. Experiments involving total RNA but not cell lysates did not show the observed drift in the control trace (data not shown).

tions needed for sophisticated applications. Our entropy-driven catalytic reaction networks are suited for scaling up to larger circuits. The modular molecular design makes synthesis of more complex components and networks with arbitrary topology straightforward. To demonstrate this, we constructed an entropy-driven catalytic analog AND gate wherein both of two catalysts are required to release output (see Text 4-S12 and Fig. 4-S11 [22]). For scaling up to large circuits, independent catalyst systems must have negligible crosstalk. The success of quantitative models that assume no crosstalk, as presented above, is encouraging; further evidence comes from a test of two independent catalyst systems operating in the same solution (see Fig. 4-S12 [22]). Finally, catalytic systems have the potential to avoid the slow-down that plagued previous attempts to construct large nucleic acid circuits

[17].

Future nucleic acid control circuits must be interfaced to molecular sensors and actuators. This may be achieved directly when the inputs and outputs are themselves nucleic acids, such as for the detection, analysis, and response to complex nucleic acid samples [9, 30] or for the control of nucleic acid nanomachines [2, 31]. Nucleic acid circuits can also respond to and control more general chemical events: In principle, release of an oligonucleotide could regulate covalent chemistry by controlling (deoxy)ribozyme activity [9] or reactant proximity [32]. Additionally, signals carried by small organics and other non-nucleic-acid molecules can be read by nucleic acid systems using aptamer domains [33, 34] and other binding interactions that can regulate toehold accessibility [35, 36]. Thus, nucleic acids could provide a general-purpose system for synthesis of embedded control circuitry within aqueous chemical systems.

Acknowledgements We thank Robert Xiaoyan Bao, Georg Seelig, David Soloveichik, Paul Rothemund, and Len Adleman for insightful discussions. There is a patent pending on this work. DYZ and AJT were supported by the UK research councils EPSRC, BBSRC, MRC and the MoD though the Bionanotechnology Interdisciplinary Research Collaboration. DYZ and EW were supported by a Caltech Grubstake Grant and NSF grants 0506468, 0622254, and 0533064. DYZ is supported by the Fannie and John Hertz Foundation.

4-S1. DNA sequence design

The DNA sequence design process was done on a domain level. Domains denoted by unmodified numbers are termed primary domains, while domains denoted by barred numbers are termed complementary domains. First, random sequences composed of only A, C, and T were generated for each of the primary domains. Since in the intended reaction pathway only primary domains are ever exposed as single-stranded regions, this choice reduces potential secondary structure [37]. Sequences for complementary specificity domains were constructed accordingly. Next, subsequences known to be problematic (4 or more G's in a row on complementary domains due to G-quadruplexing, more than 4 A's in a row caus-

ing synthesis difficulties, etc.) were altered by hand. The remaining sequences were then concatenated as appropriate to form the DNA strands. These were folded alone and pairwise using the mFold web-server to determine possible spurious bindings [38]. Some bases at problematic subsequences were then changed by hand to G in the primary domains (and propagated elsewhere as appropriate for domain identity), to minimize self-folding and pairwise-folding energies. Finally, the strands were checked again on mFold to ensure minimal spurious interactions [38].

4-S2. Materials and methods

Substrate purification. Substrate and reporter complexes were manually purified to ensure proper stoichiometry and to improve purity. Sources of substrate impurity include synthesis errors and truncations, partially-formed complexes due to imperfect stoichiometry, and dimerization. Strands for each sample were prepared with nominally correct stoichiometry at 20 μM and annealed. For all substrate complexes except the autocatalyst substrate, the fuel strand was then added, which triggers many poorly-formed substrates to decay into products that can be removed by gel purification. (For the autocatalyst, addition of the fuel strand would have initiated the exponential chain reaction, so the autocatalyst substrate was purified without addition of the fuel strand.) The samples were then run on 12% non-denaturing polyacrylamide gel electrophoresis (PAGE) gel at 180V for 6 hours. The proper bands were cut out and eluted in TE/Mg^{2+} buffer for 2 days. Typical prep sizes ranged from 5 nmol to 10 nmol, and typical elution volume was 2 ml. Typical yields ranged from 40% to 60%. Purified complexes were quantitated by measurement of absorbance at 260 nm, using extinction coefficients for single- and double-stranded DNA predicted by nearest-neighbor models [39].

All annealing processes were performed with an Eppendorf Mastercycler Gradient thermocycler. The samples were brought down from 95 °C to 20 °C at a constant rate over the course of 90 minutes.

DNA oligonucleotides. DNA oligonucleotides used in this study were purchased from Integrated DNA Technologies(IDT), with HPLC purification. Where applicable, flu-

orophores were attached by IDT as well.

Buffer conditions. The buffer for all experiments was TE (10 mM Tris · HCl pH balanced to 8.0, 1mM EDTA), purchased in 100x stock from Sigma-Aldrich (catalog number T9285), with 12.5 mM $MgCl_2$ added.

Gel electrophoresis. Non-denaturing PAGE was run on 12% acrylamide (19:1 acrylamide:bis), diluted from 40% acrylamide stock purchased from Ambion (catalog number AM9022). ND loading dye containing XCFF in 50% glycerol was added in 0.2x stochiometry to all samples. Gels were run at 25 °C using a Novex chamber with external temperature bath. Gels were stained with Sybr-Gold stain, purchased from Invitrogen (catalog number S-11494), and scanned with a Bio-Rad Molecular Imager. Formation gels shown in Fig. 4-1C and Fig. 4-2 were run at 180V for 1 hour.

Total RNA and cell lysate. In the experiment described in Fig. 4-4B, inset 2, mouse liver total RNA and active rabbit reticulocyte lysate were used. They were both purchased from Ambion (catalog numbers AM7810 and AM1200), as part of their in vitro translation kit. The lysate included exogeneously introduced RNAse inhibitor enzymes; we do not expect this to significantly affect the results of the experiment described in Fig. 4-4B inset 2.

Spectrofluorimetry studies. Spectrofluorimetry studies were done with a commercial SPEX Fluorolog-3 from Horiba. Cuvettes used were 119-004F synthetic quartz cells purchased from Hellma, with total volume 1.6 ml. For studies observing behavior of the TET fluorophore, excitation were at 524 nm, while emissions were at 541 nm. For studies observing behavior of the ROX fluorophore, excitation were at 588 nm, while emissions were at 602 nm. For studies observing behavior of the TAMRA fluorophore, excitation were at 557 nm, while emissions were at 580 nm. Slit size used are 2 nm for both excitation and emission monochromators for net reaction studies, and 3 nm for individual rate measurements. All experiments were done with integration time of 3 seconds for every 30 second time-point.

Prior to each experiment, all cuvettes were cleaned thoroughly: each cuvette was washed 15 times in distilled water, once in 70% ethanol, another 5 times in distilled water, and finally

once more in 70% ethanol.

Fluorescence normalization. All fluorescence experiments show fluorescence values normalized to approximately 1 a.u. = 10 nM. Simulation traces (dotted lines) are offset vertically to correspond to quenched fluorophore baselines. Data traces within a single figure are normalized using the same scaling factor, which was determined by best-fit to simulation traces. Data traces across different figures possess different scaling factors due to differences in fluorescence reporter, lamp luminosity, and substrate concentrations. Time $t = 0$ signals the beginning of the reaction, triggered by the addition of the last necessary reagent (usually the substrate).

Inactivity of carrier strands. In the course of testing the catalyst system and its derivatives, some reactions required very small quantities of certain DNA species. For example, in Fig. 4-3B (inset), 1 pM of $C0$ in 1.5 ml of solution = 1.5 fmol of DNA. We have experimentally observed that DNA sticks non-specifically to pipette tips, so that direct serial dilutions lead to stocks much more dilute than expected. Unfortunately, this loss is not very consistent, so we could not compensate for tip loss with additional reagent. Instead, we worked around this issue by introducing into all dilute stocks (1 μM and below) a non-reactive 20 nt poly-T "carrier" strand, at a concentration of 1 μM. Since pipette tip loss is non-specific, the majority of DNA loss would be of the carrier strand, so that serially diluted stocks are only slightly more dilute than expected. It is of interest to verify that the carrier strand does not affect kinetics. Thus, we compared performance of the catalyst at concentration ranges where pipette loss is not substantial (see Fig. 4-S1). Presence of the carrier strand appears to have very little, if any, effect on the kinetics of the catalyst, at 100x excess.

4-S3. Directly-labelled cross-catalyst

This section refers to our earlier attempts at constructing a cross-catalytic network without using the indirect reporting scheme presented in the main paper. Instead, we directly tagged the fuel strand $F3$ with the HEX (Hexachlorofluorescein) fluorophore. As the catalyzed reaction occurs, the fuel strand $F3$ is co-localized with linker strand $LB3$ in

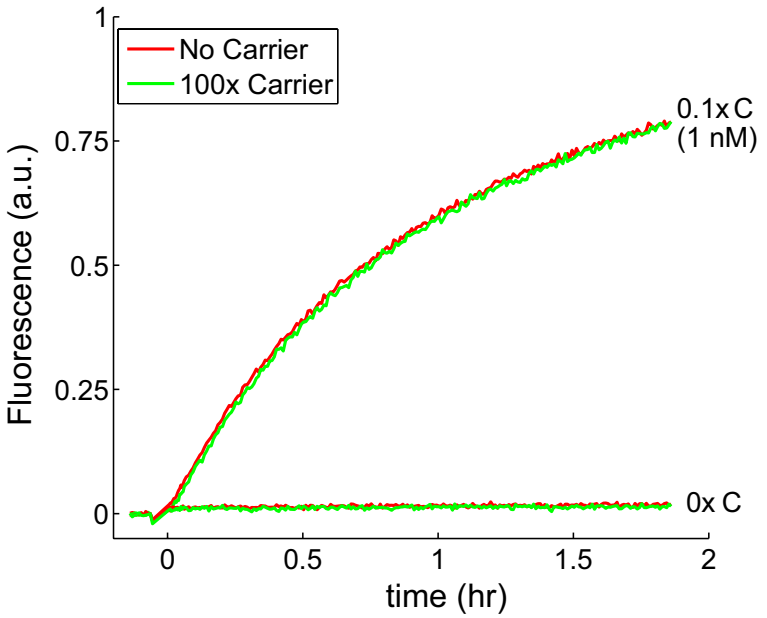


FIG. 4-S1: Non-interaction of poly-T carrier strand. Substrate concentration $[S] = 10$ nM = 1x (15 pmol at 1.5 ml volume); fuel concentration $[F] = 13$ nM; 25 °C. Presence of 100x = 1 μM (1.5 nmol for this reaction) of carrier had little effect on kinetics of either catalyzed or uncatalyzed reaction rate.

the waste product, and the HEX fluorophore is quenched due to its close contact with a guanine (G) base [40]. The sequences of the involved domains are given in Table 4-S1.

Note that the kinetic behavior of cross-catalytic circuit differs substantially from that presented in section S11. First, the reaction is substantially slower, despite being at higher concentrations. Second, the reaction does not quickly go to completion, as the autocatalyst and cross-catalyst circuits presented earlier. We interpret this to be evidence for interaction between the HEX fluorophore and DNA.

4-S4. Reporter complex characterizations

In the course of this work, we observe kinetics using independent reporter complexes OR and SR; this approach was chosen (rather than direct labelling of strands in the catalyst system) to decouple the thermodynamic effects of fluorophore-quencher binding from the catalytic pathway. Both OR and SR initially contain a 20 bp duplex and a 7 nt toehold domain that uniquely binds their respective targets (OB and SB). Each possesses a different fluorophore and quencher pair (TET and Iowa Black Fluorescence Quencher (FQ) for SR; ROX and Iowa Black Red Quencher (RQ) for OR). The reactions are assumed to be

Dom.	Sequence	Length
21	5'- TATTCC -3'	6
22	5'- GCTA -3'	4
23	5'- GTCA -3'	4
24	5'- TACCAA -3'	6
25	5'- CATCG -3'	5
26	5'- ACTACAC -3'	7
27	5'- CTCAG -3'	5
28	5'- CTCAATC -3'	7
29	5'- TACTCTACAC -3'	10
30	5'- CAAATCCTCA -3'	10

Table 4-S1: Domain sequences of HEX-labelled systems

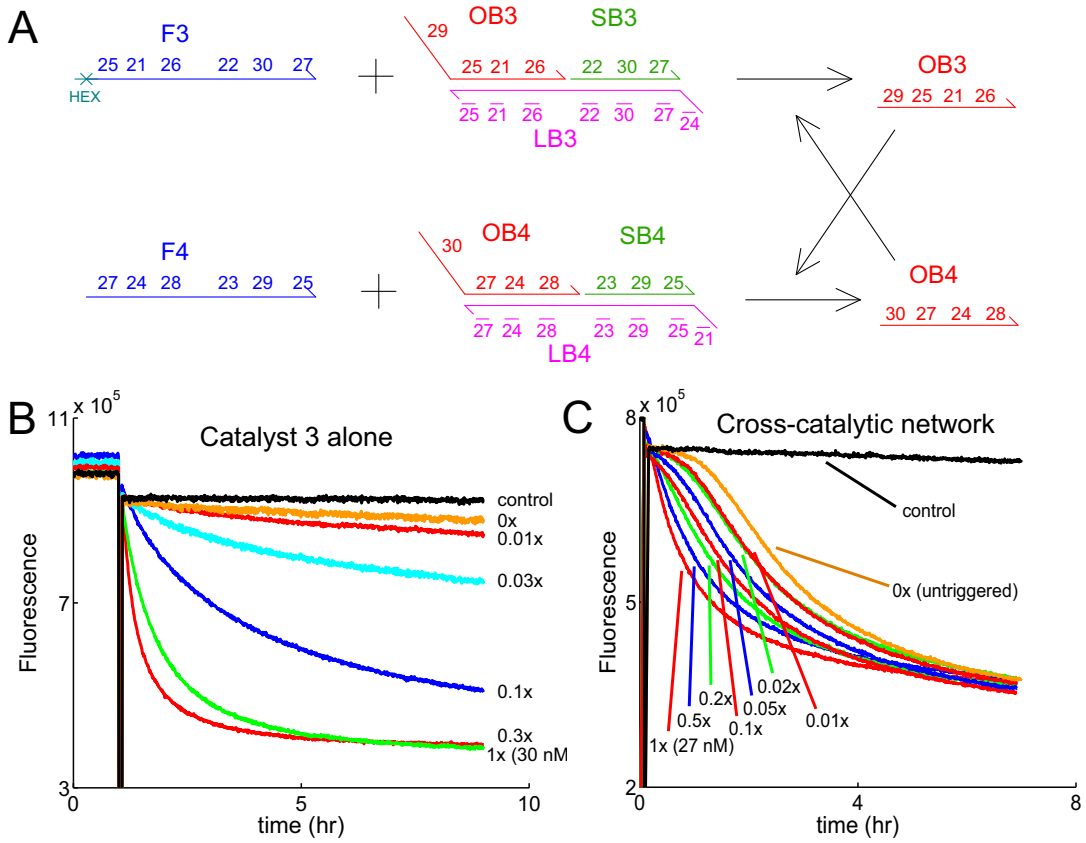


FIG. 4-S2: Directly-labelled cross-catalyst circuit **(A)** Cross-catalyst schematic. HEX represents the hexachlorofluorescein fluorophore, and is quenched by the 3' most base of the the $LB3$ is a guanine (G). **(B)** Test of the F3 + S3 reaction in isolation. $[F3] = [S3] = 30$ nM, $[F4] = [S4] = 0$. Various amounts of catalyst $OB4$ were added at t $\approx$ 1 hour. The catalyzed reaction is substantially slower than that shown in Fig. 4-1E, despite being at higher concentration. Additionally, the control reaction (which lacked substrate $S3$) showed some decrease of signal over the observed time; this is suspected to be fluorophore bleaching. These were the earliest experiments we ran, and technical proficiency for these experiments were not as high as for other results. Normalization was not performed for these reactions because it is not clear where endpoints were. **(C)** Cross-catalyst series. $[F3] = [F4] = [S3] = [S4] = 27$ nM. Only catalyst $OB3$ was added at t $\approx$ 0.

non-reversible:

$$SB + SR \overset{k_{TET}}{\rightarrow} Fluorescence \tag{1}$$

$$OB + OR \overset{k_{ROX}}{\rightarrow} Fluorescence \tag{2}$$

Displacement rate constants of the two reporter complexes OR and SR were measured via fluorescence at initial reactant concentrations of 1 nM ($\frac{1}{30}$ the concentration used in catalyst experiments) to be $k_{TET} = 8 \cdot 10^5 M^{-1}s^{-1}$ and $k_{ROX} = 4 \cdot 10^5 M^{-1}s^{-1}$ (see Fig. 4-S3). The observed difference in displacement rate constants may be due to either the thermodynamics of the fluorophore/quencher pairs, differences in binding strength of the toehold domains, or secondary structure differences in SB and OB.

4-S5. Catalyst modelling and characterization

As described in the main paper, we are able to predict the net behavior of the catalyst system using a reduced model of the catalytic pathway. The reactions of three-step process are redisplayed below.

$$S + F \overset{k_0}{\rightarrow} OB + SB + W \tag{3}$$

$$S + C \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} I3 + SB \tag{4}$$

$$I3 + F \overset{k_2}{\rightarrow} I5 + OB \tag{5}$$

$$I5 \underset{k_{-3}}{\overset{k_3}{\rightleftharpoons}} C + W \tag{6}$$

Using symmetry and our understanding of DNA binding thermodynamics, we are able to approximate two of the parameters: First, the spurious re-association rate of C and W is initiated by the same external 6 nt $\bar{5}$ domain as the correct association of C to substrate S, and consequently the rate constants should be similar. Thus, we assume for our simulations that $k_{-3} = k_1$. Second, the back-reaction of $I3$ and SB to re-form S and C is initiated by the same internal domain $\bar{5}$ (of length 4 nt) as the correct association of F to intermediate

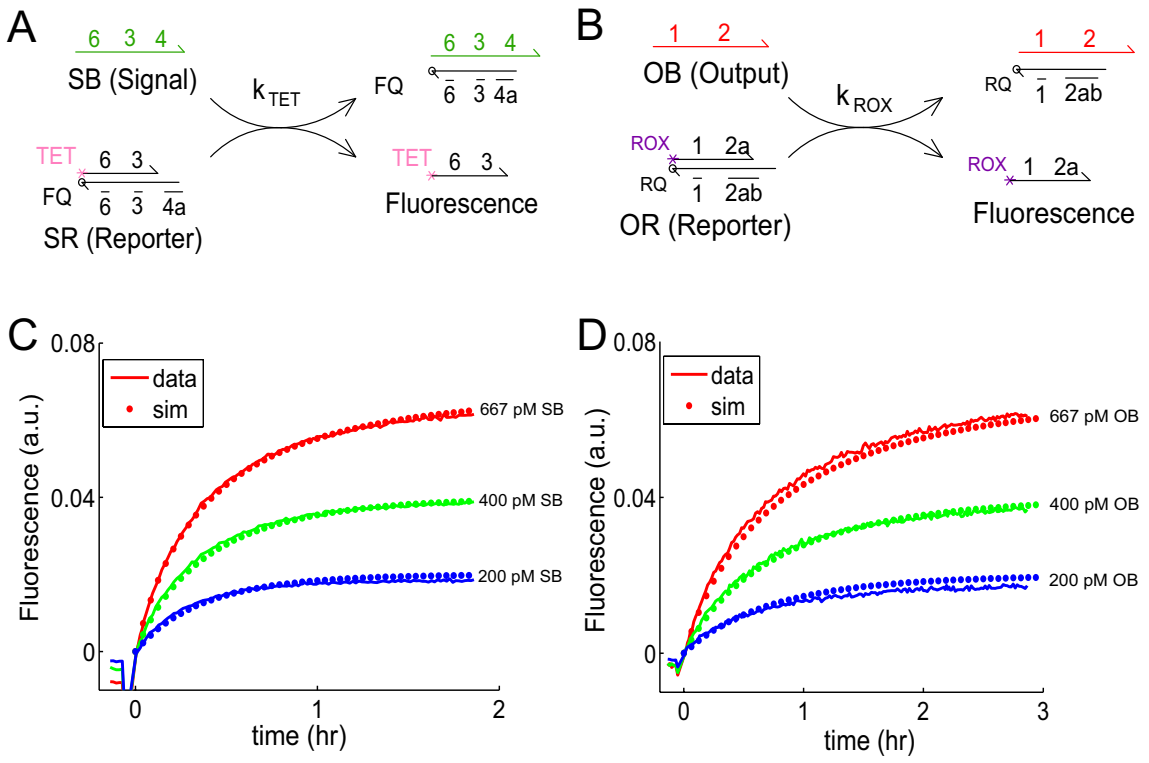


FIG. 4-S3: Reference experiments. **(A)** Schematic of signal (SB) reporter using TET and Iowa Black Fluorescence Quencher (FQ). **(B)** Schematic of output (OB) reporter using ROX and Iowa Black Red Quencher (RQ). **(C)** Displacement rate of strand SF by SB at 25 °C. Reporter complex SR was present in solution at 1 nM at t=0, and various amounts of SB were added at t ≈ 0. Dotted lines show simulation traces of a second-order displacement reaction with rate constant $k_{TET} = 8 \cdot 10^5 M^{-1} s^{-1}$. **(D)** Displacement rate of strand OF by OB at 25 °C. Reporter complex OR was present in solution at 1 nM at t=0, and various amounts of OB were added at t ≈ 0. Dotted lines show simulation traces of a second-order displacement reaction with rate constant $k_{ROX} = 4 \cdot 10^5 M^{-1} s^{-1}$.

$I3$, so we assume that $k_{-1} = k_2$.

Using k_{TET} and k_{ROX} rate constants from the previous section, reaction rate constants k_1, k_2, and k_0 are directly measured to be $6.5 \cdot 10^5$, $4.2 \cdot 10^5$, and $2.3 \cdot 10^1 M^{-1} s^{-1}$, respectively (see Fig. 4-S4). Note that for these experiments, we can effectively ignore the k_{-1} rate of reverse reaction $SB + I3$, because SB was consumed by reporter complex SR (simulations showed no visible difference when k_{-1} was modeled). The last rate constant k_3 is difficult to measure because it is first-order, and we could not slow it down to the time-scale where our spectrofluorimeter readings are meaningful. Thus, it was fitted using the results of the net kinetics of the catalytic system to be $4 \cdot 10^{-3} s^{-1}$. Fig. 4-1E is redisplayed here as Fig. 4-S4D.

4-S6. Estimates of Entropy and Free Energy

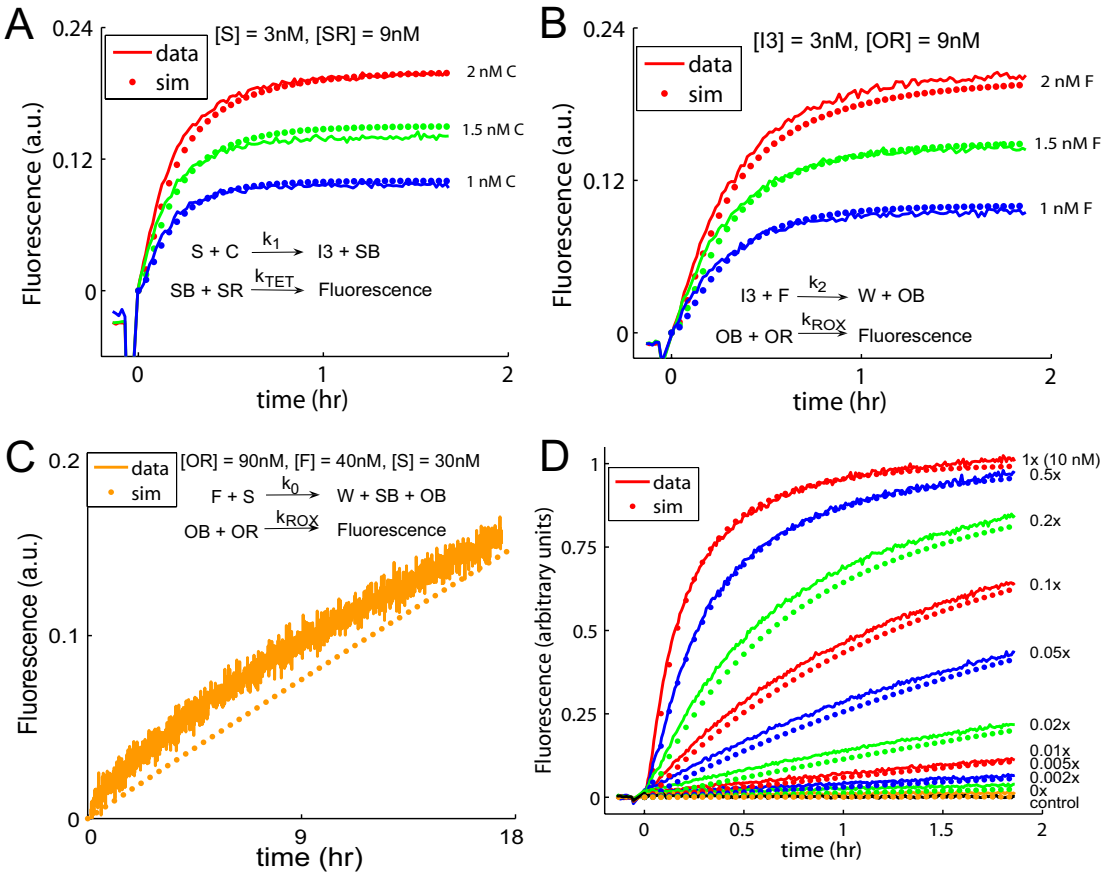


FIG. 4-S4: Intermediate rate measurements. **(A)** Measurement of catalyst binding rate k_1. All traces contain 3 nM S initially, and different amounts of C were added at $t \approx 0$. Reporter SR was present at 9 nM concentration. Dotted lines show simulation traces modelling reactions (1) and (4), assuming rate constant $k_1 = 6.5 \cdot 10^5 M^{-1}s^{-1}$ **(B)** Measurement of fuel binding rate k_2. All traces contain 3 nM pre-prepared $I3$ initially, and different amounts of F were added at $t \approx 0$. Reporter OR was present at 9 nM concentration. Dotted lines show simulation traces modelling reactions (2) and (5), assuming rate constant $k_2 = 4.2 \cdot 10^5 M^{-1}s^{-1}$ **(C)** Measurement of uncatalyzed reaction rate k_0. $[S] = 30$ nM, $[F] = 40$ nM, and no catalyst was present. Reporter OR was present at 90 nM concentration. Dotted lines show simulation traces modelling reactions (2) and (3), assuming rate constant $k_0 = 2.3 \cdot 10^1 M^{-1}s^{-1}$ **(D)** Data vs. simulation using measured rates, repeated from main text. Rate constant $k_3 = 4 \cdot 10^{-3}s^{-1}$ was fitted.

We consider the net reaction

$$S + F \rightleftharpoons OB + SB + W$$

The free energy change for this reaction, in dilute solutions, is

$$\Delta G = \Delta G^\circ_{OB} + \Delta G^\circ_{SB} + \Delta G^\circ_W - \Delta G^\circ_S - \Delta G^\circ_F + RT \ln Q \stackrel{\text{def}}{=} \Delta G^\circ_{net} + RT \ln Q$$

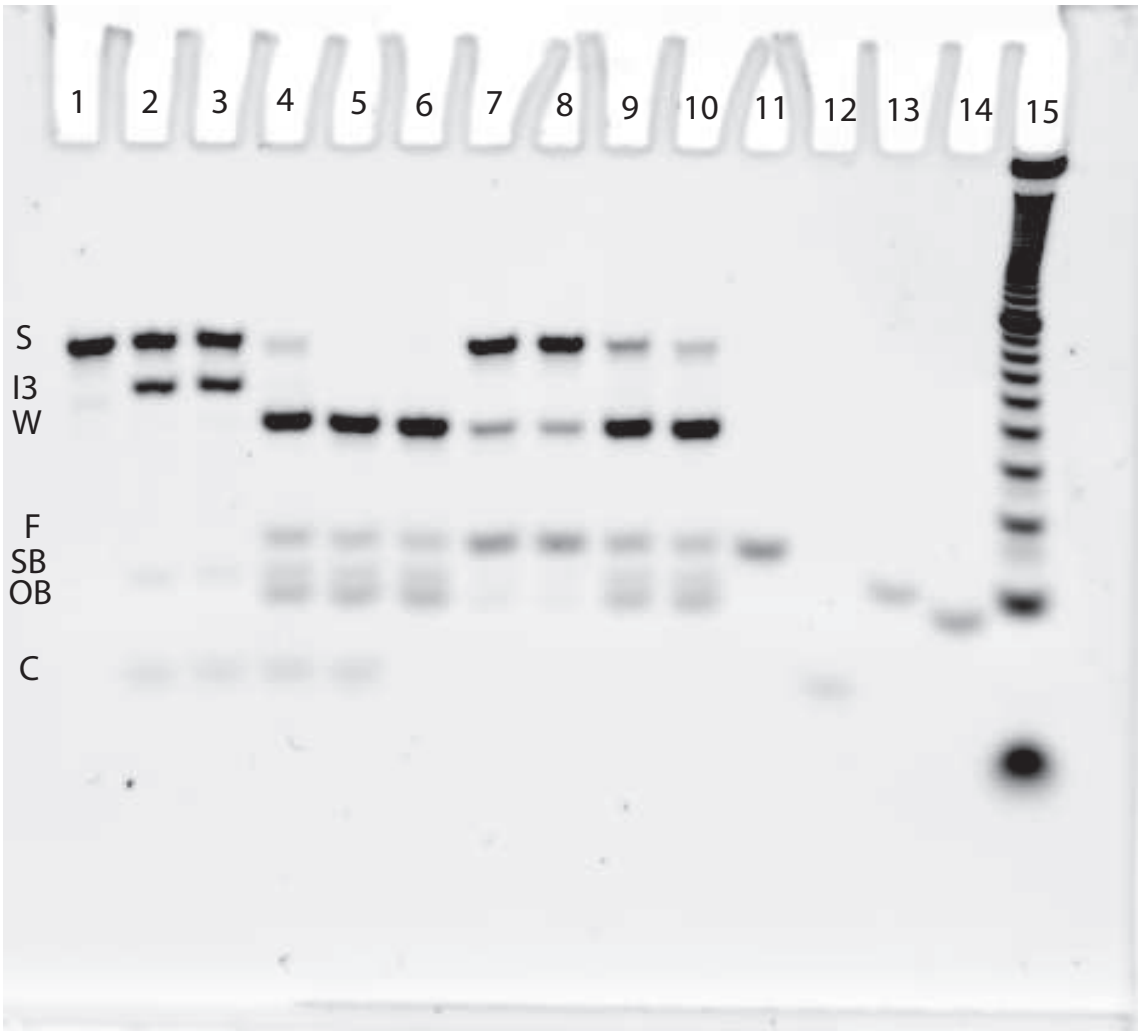


FIG. 4-S5: Full formation gel including control lanes. 25 °C, 1x = 200 nM, 12% non-denaturing (native) PAGE. This is the full version of the cropped gel in Fig. 4-1C. Note that Sybrgold stains single-stranded DNA much less efficiently than double-stranded DNA, so that relative brightness between different bands should not be compared.
L1: S [1x]; L2: S [1x] + C [1x], 30 min; L3: (S [1x] + C [1x]); L4: (S [1x] + C [1x]) + F [1x], 30 min; L5: (S [1x] + C [1x] + F [1x]); L6: (S [1x] + F [1x]); L7: S [1x] + F [1x], 30 min; L8: S [1x] + F [1x], 180 min; L9: S [1x] + F [1x] + C [0.1x], 30 min; L10: S [1x] + F [1x] + C [0.1x], 180 min; L11: F [1x]; L12: C [1x]; L13: SB [1x]; L14: OB [1x]; L15: 10 nt duplex ladder.
Parentheses denote that the included species were annealed together.

where $Q = ([OB]/c^\circ \cdot [SB]/c^\circ \cdot [W]/c^\circ)/([S]/c^\circ \cdot [F]/c^\circ)$ is the reaction quotient relative to standard conditions and ΔG°_X is the standard free energy of species X at standard conditions, which here specify our TE buffer with 12.5 mM magnesium, 25 °C, and $c^\circ = 1$ M.

The free energy change (the driving force for the reaction) decreases as concentrations change during the course of the reaction; once equilibrium is achieved, $Q =$

$\exp\{-\Delta G^\circ_{net}/RT\}$ and $\Delta G = 0$. If the standard free energy change $\Delta G^\circ_{net} \approx 0$, as we expect for the reaction with the full-length fuel strand if the standard free energy is dominated by base pairing, then the driving force at any moment is just $RT \ln Q$. As a somewhat arbitrary reference point, we consider the time at which half the substrate has been depleted. For the reaction in Fig. 4-2, this occurs when $[S] = [F] = [SB] = [OB] = [W] = c = 100$ nM, $Q = 10^{-7}$, and $RT \ln Q = RT \ln c = -9.6$ kcal/mol. For the reaction in Fig. 4-1E, $c = 5$ nM and $RT \ln c = -11.4$ kcal/mol.

The free energy difference between the substrate S and the maximally truncated waste product W was approximated using the mFold server [38] using DNA parameters for 25 °C, with salt conditions being 10 mM Na^+ and 12.5 mM Mg^{2+}. Taking into consideration the 8 base pair stacks, external loops and dangles (due to the 1 domain in S, and the 3' overhang on the LB strand on the truncated waste product W), and an initiation entropy of 6.4 cal/mol/K per association, the predicted standard free energy change ΔG°_{net} for the (unfavorable) forward reaction is +11.7 kcal/mol.

According to these estimates, truncating the fuel strand F by 8 bases should disfavor the forward reaction enough that the equilibrium distribution possesses substrate S in excess of waste W. However, the experiments described in Fig. 4-2 show waste in excess of substrate after 3 hours. This suggests that the estimate for ΔG°_{net} is too large; a value closer to +9 would be more compatible with the experiments.

4-S7. Catalyst robustness to temperature and salt conditions

As mentioned in the main text, we expect the entropy-driven catalytic reaction design to be somewhat more robust to changes in environmental conditions, such as temperature and salt concentrations. Salt conditions affect the free energy of hybridization [41], but since there is no net gain or loss of base-pairs in a net reaction cycle, the equilibrium should not be significantly affected. The strength of the binding of the toehold domains still depend on the free energy of the base pairs formed, so salt concentration will affect the kinetics of the catalyzed pathway. However, the catalyst should qualitatively function across the range of salt concentrations, as long as the toehold domains still are able to co-localize the relevant

strands. In Fig. 4-S6A, we see that catalytic function is preserved across a broad range of salt buffers, but degrades for sufficiently low salt concentrations.

Temperature affects the equilibrium of a reaction only through the enthalpic change (ΔH°). When the magnitude of the enthalpic change is small ($\Delta H^\circ \approx 0$) as it is for the entropy-driven system, the effects of temperature have little effect on the reaction equilibrium – although, again, they affect the kinetics of toehold-mediated processes. In Fig. 4-S6B, the catalyst is shown to function across a 25 degree range of temperatures, from 12 °C to 37 °C.

4-S8. Details of quadratic feedforward circuit and autocatalyst

Domain sequences. The domains involved in the quadratic feedforward circuit (Fig. 4-3) and the autocatalytic reaction (Fig. 4-4) are shown below in Table 4-S2. As mentioned in the main text, there is a significant amount of domain redundancy, because the initial catalyst, quadratic circuit, and autocatalyst circuit were designed simultaneously with the goal of minimizing the number of changes between designs. Again, we stress that in designing a catalytic reaction system in isolation, the domains may be completely independent (see section S10).

Domains $2a$, $4b$, and $9b$ are identical to each other, and are also referred to as x in the cross-catalytic schematic (see Fig. 4-S10). Domains $2b$, 5, and 10 are identical to each other, and are also referred to as y in the cross-catalytic schematic.

Quadratic feedforward circuit fits. For the feedforward circuit, the model presented in the main paper is expanded to include one separate set of equations for each layer:

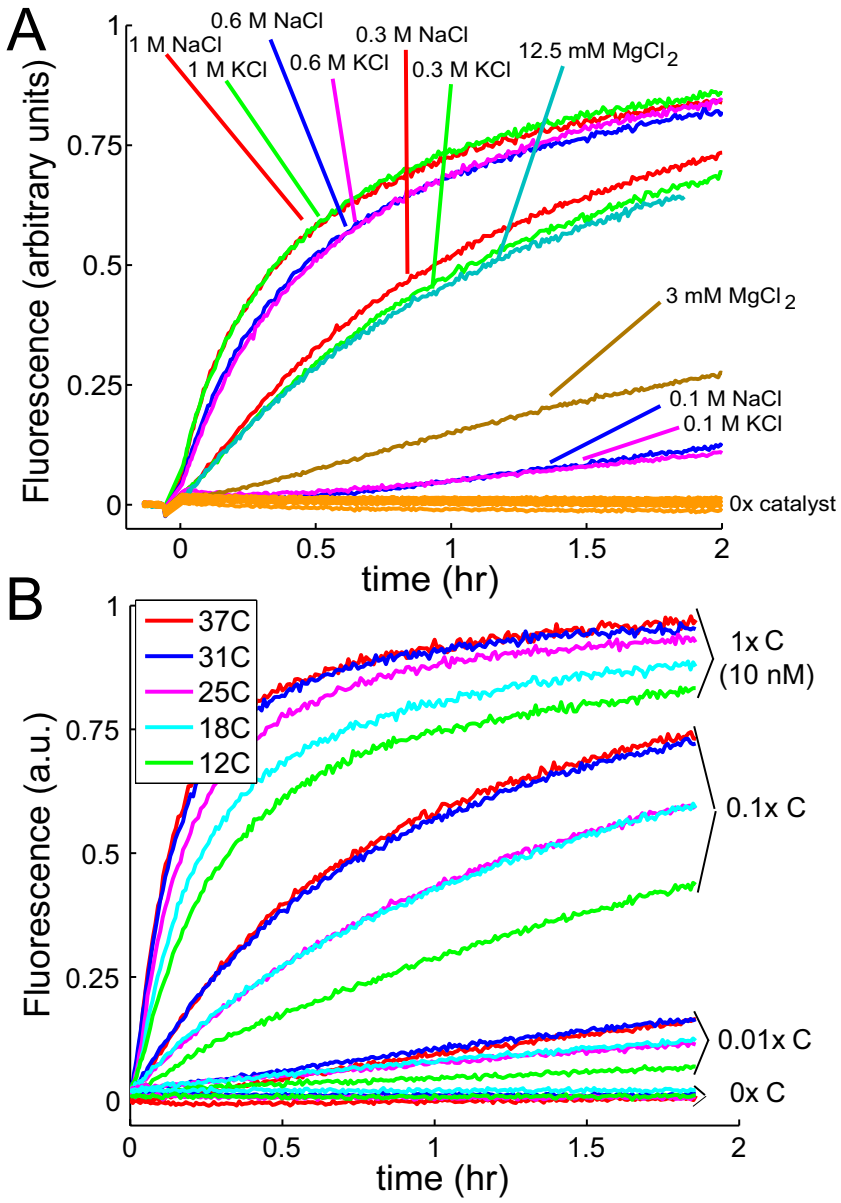


FIG. 4-S6: Robustness of entropy-driven catalytic reaction shown in Fig. 4-1. **(A)** Salt condition robustness. The same system as presented in Fig. 4-1 was tested here in TE supplemented with various different salt concentrations. $[S] = 10$ nM and $[F] = 13$nM. The orange traces at the bottom (controls run for every salt condition tested) show $[C] = 0$, while all other traces have $[C] = 1$ nM $= 0.1$x. Reactions were run at 25 °C. Catalysis functions with fast kinetics over a broad range of salt conditions. All traces shown normalized to the same scaling factor. **(B)** Temperature robustness. Catalysis functions qualitatively similar across temperatures from 12 °C to 37 °C. Reactions were run in TE with 12.5 mM $MgCl_2$. Note that the 18 °C and 25 °C traces show very similar kinetics for 0.1x C and 0.01x C, as do the 31 °C and 37 °C traces. To account for temperature-dependent fluorescence, traces were normalized individually with separate experiments (data not shown) that measured the fluorescence of the ROX fluorophore at different temperatures. We do not have an explanation for the near-superposition of some traces run under different conditions.

$$S0 + F0 \overset{k_{00}}{\rightarrow} OB0 + SB0 + W0 \tag{7}$$

$$S0 + C0 \underset{k_{-10}}{\overset{k_{10}}{\rightleftharpoons}} I30 + SB0 \tag{8}$$

$$I30 + F0 \overset{k_{20}}{\rightarrow} I50 + OB0 \tag{9}$$

$$I50 \underset{k_{-30}}{\overset{k_{30}}{\rightleftharpoons}} C0 + W0 \tag{10}$$

$$S1 + F1 \overset{k_{01}}{\rightarrow} OB1 + SB1 + W1 \tag{11}$$

$$S1 + OB0 \underset{k_{-11}}{\overset{k_{11}}{\rightleftharpoons}} I31 + SB1 \tag{12}$$

$$I31 + F1 \overset{k_{21}}{\rightarrow} I51 + OB1 \tag{13}$$

$$I51 \underset{k_{-31}}{\overset{k_{31}}{\rightleftharpoons}} C1 + W1 \tag{14}$$

$$\tag{15}$$

As mentioned in the main text, the downstream layer catalyst system is identical to that presented in Fig. 4-1C, and thus we use the same rate parameters for it as used in Fig. 4-1E: $k_{01} = 2.3 \cdot 10^{1} M^{-1} s^{-1}$, $k_{11} = k_{-31} = 6.5 \cdot 10^{5} M^{-1} s^{-1}$, $k_{-11} = k_{21} = 4.2 \cdot 10^{5} M^{-1} s^{-1}$, $k_{31} = 4 \cdot 10^{-3} s^{-1}$. The remaining parameters for the upstream catalyst system are fitted to the quadratic series data, with the same constraints $k_{10} = k_{-30}$ and $k_{-10} = k_{20}$. The dotted traces displayed in Fig. 4-S7C show $k_{10} = k_{-30} = 5.0 \cdot 10^{5} M^{-1} s^{-1}$, $k_{11} = k_{21} = 1.9 \cdot 10^{5} M^{-1} s^{-1}$, and $k_{30} = 8 \cdot 10^{-3} s^{-1}$. Note that the rates are quantitatively similar (though somewhat slower) to the rates of the analogous reactions in the upstream system, which supports our understanding of the behavior of the circuit.

As seen in the gel in Fig. 4-1C, some impurity fraction of the substrate S reacts very quickly to fuel F. Evidence for this is also seen in Fig. 4-1E, where the fluorescence value of the yellow "0x" trace is slightly higher after the addition of fuel F at $t \approx 0$. For the quadratic feedforward circuit, a small constant amount of $OB0$ released from the upstream system would be converted into a small constant production rate of $OB1$. We fitted this initial concentration $[OB0]$ to be 45 pM, for a total of 4 additional fitted parameters.

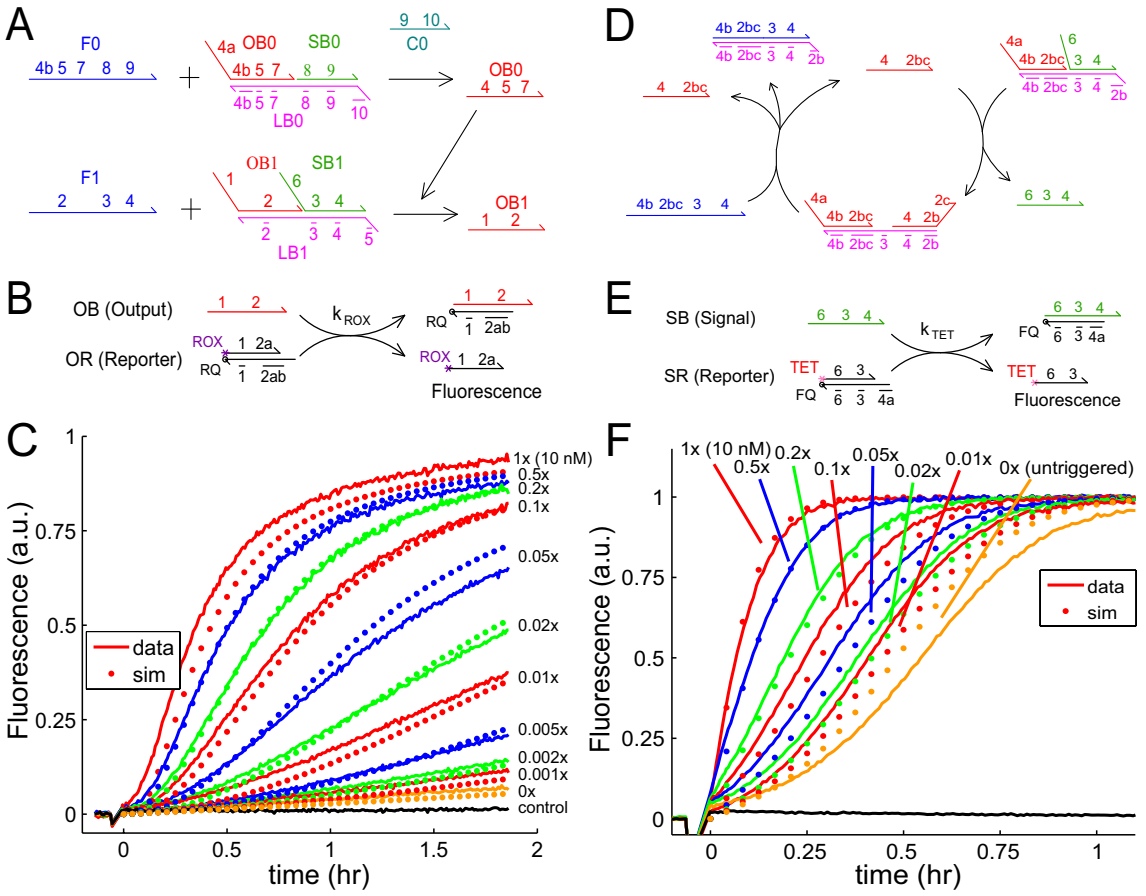


FIG. 4-S7: Fitting the feedforward circuit and the autocatalytic reaction. **(A)** Schematic of the feedforward circuit. **(B)** Schematic of the reporter OR used for the feedforward circuit (same as displayed in Fig. 4-1D). **(C)** Fits to the quadratic circuit using fitted parameters described in text. **(D)** Schematic of the autocatalytic reaction. **(E)** Schematic of the reporter SR used for the autocatalytic reaction and the cross-catalytic circuit. **(F)** Fits to the autocatalytic reaction using the fitted parameter for the initial $[OB]$, as described in the text.

Dom.	Sequence	Length
1	5'- CTTTCCTACA -3'	10
2a (=x)	5'- CCTACG -3'	6
2b (=y)	5'- TCTCCA -3'	6
2c	5'- ACTAACTTACGG -3'	12
3	5'- CCCT -3'	4
4a	5'- CATTCAATAC -3'	10
4b (=x)	5'- CCTACG -3'	6
5 (=y)	5'- TCTCCA -3'	6
6	5'- CCACATACATCATATT -3'	16
7	5'- TACTTATTAGCC -3'	12
8	5'- GACA -3'	4
9a	5'- CTACTTTCAC -3'	10
9b (=x)	5'- CCTACG -3'	6
10 (=y)	5'- TCTCCA -3'	6

Table 4-S2: Domain sequences

Autocatalytic feedback circuit fits. For the autocatalytic system, we use reporter complex SR to measure the progress of the reaction, because we do not wish to damp the exponentiation by reacting OB with OR. A schematic for the reaction is shown in Fig. 4-S7E. The reaction set we use to model the behavior of the autocatalyst is shown below:

$$SB + SR \overset{k_{TET}}{\rightarrow} Fluorescence \tag{16}$$

$$S + F \overset{k_0}{\rightarrow} OB + SB + W \tag{17}$$

$$S + OB \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} I3 + SB \tag{18}$$

$$I3 + F \overset{k_2}{\rightarrow} I5 + OB \tag{19}$$

$$I5 \underset{k_{-3}}{\overset{k_3}{\rightleftharpoons}} OB + W \tag{20}$$

Parameters k_{TET}, k_0, k_1, k_{-1}, k_2, and k_{-3} again are the same as measured previously. Parameter $k_3 = 4 \cdot 10^{-3} s^{-1}$ is the same as fitted to the catalyst data in Fig. 4-1E. The only new parameter fitted is the impurity concentration, the initial concentration of $[OB] =$ 460 pM. Purification of the autocatalyst substrate was significantly more difficult technically

than that of the catalyst substrate, as evidenced by different inferred initial $[OB]$ for different purified samples.

4-S9. Repeatability of fluorescence experiments

Several factors reduced the repeatability of fluorescence experiments: First, the spectrofluorimeter luminosity output differs from lamp bulb to lamp bulb and luminosity tends to decrease as any particular lamp bulb ages. Second, different preparations of purified substrate complexes S, though nominally calibrated to the same concentration, in practice differed in purity. Third, fluorophores tend to bleach, and thus older stocks tend to give lower fluorescence readings for the same concentration. Finally, our Eppendorf pipettes are high precision but low accuracy; thus using two different pipettes to measure the same volume would often yield different pipetted quantities.

The effects of the above on repeatability is shown in Fig. 4-S8: Fig. 4-S8A shows repeatability of experiments before implementing measures to combat any of the above (traces agree to within 15%), while Fig. 4-S8B shows repeatability of experiments after implementing measures (traces agree to within 3%). In summary, the following was done for all fluorescence experiments displayed in the main text:

1. All traces within a figure were performed in a single sitting, one right after another.

2. All traces within a figure used the same stocks of all purified samples, including reporter complexes and substrate complexes.

3. All traces within a figure used the exact same pipette for each quantity measured (i.e. dedicated pipette for 6 μl, another dedicated pipette for 15 μl, etc).

4-S10. Catalyst system with independent input/output sequences

In the example system presented in Fig. 4-1 of the main text, domains $2b$ and 5 are identical, so strands C and OB are not independent. However, this is not essential; the design was chosen to minimize the number of changes needed to explain the autocatalyst design. A design in which OB and C possess independent sequences is demonstrated here

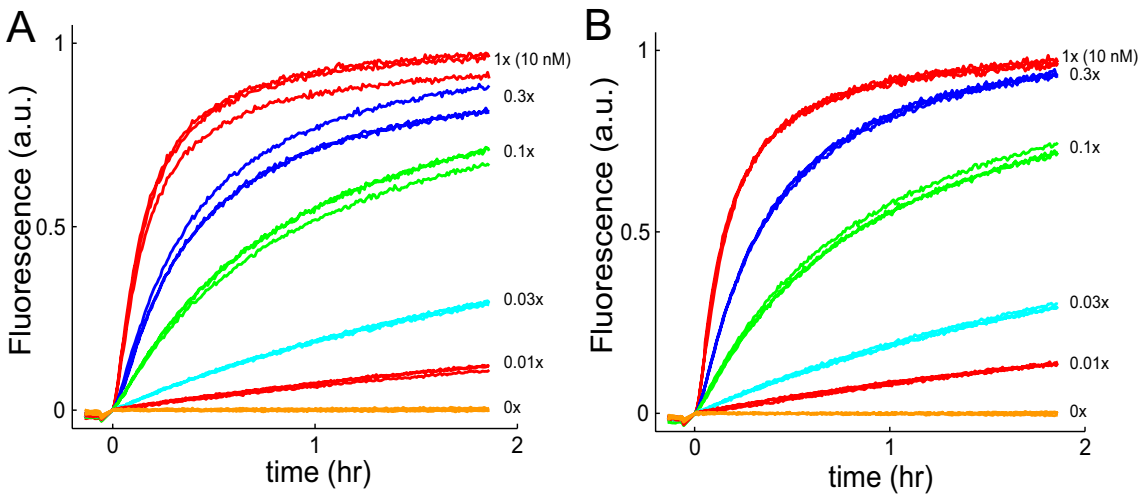


FIG. 4-S8: Repeatability of fluorescence traces. **(A)** Repeatability of experiments before optimization of protocol. The traces are of the catalyst system shown in Fig. 4-1. **(B)** Repeatability of experiments after optimization.

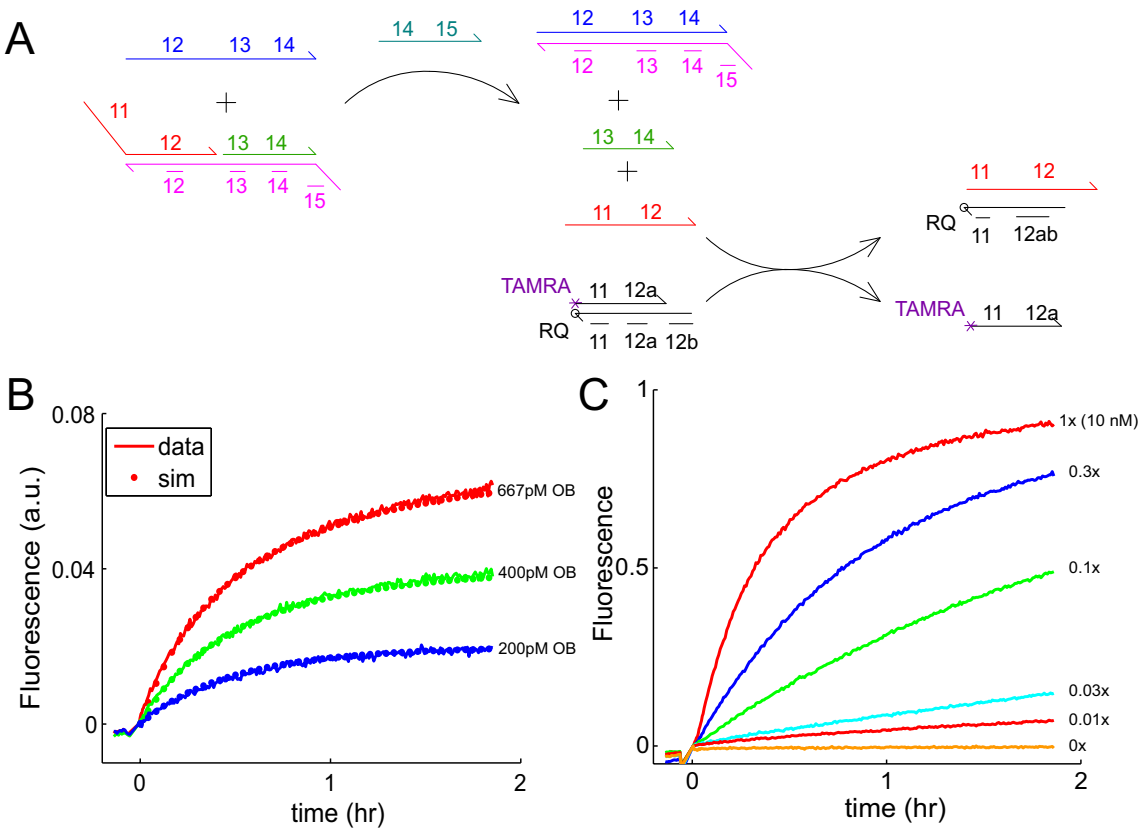


FIG. 4-S9: Catalyst system with independent input/output sequences. **(A)** Schematic of catalyst system, including fluorescence reporter. Strands are labelled by function as shown in Fig. 4-1A. **(B)** Characterization of reporter. Reporter complex was present in 1 nM concentration; various amounts of output were added. Dotted lines show simulation traces (normalized) assuming second-order displacement rate $k_{TAMRA} = 6 * 10^5 M^{-1} s^{-1}$ **(C)** Catalytic series at $[S] = 10$ nM $= 1$x, $[F] = 13$ nM $= 1.3$x.

(see Fig. 4-S9). The sequences of the system are shown in Table 4-S3. This catalyst system functions almost identically as the one presented in Fig. 4-1A.

Strand	Length	Domains	Sequence
IndCat-F	44	12 13 14 15	5'- ACCACATCAA TCTCGATCCAGTAC ACCT CTTCACGAACATTTCA -3'
IndCat-LB	50	$\bar{1}6$ $\bar{1}5$ $\bar{1}4$ $\bar{1}3$ $\bar{1}2$	5'- TGGCTA TGAAATGTTCGTGAAG AGGT GTACTGGATCGAGA TTGATGTGGT -3'
IndCat-SB	20	14 15	5'- ACCT CTTCACGAACATTTCA -3'
IndCat-OB	34	11 12 13	5'- ACCTAATAGC ACCACATCAA TCTCGATCCAGTAC -3'
IndCat-C	22	15 16	5'- CTTCACGAACATTTCA TAGCCA -3'
IndCat-OF2	20	F 11 12	5'- /TAMRA/ ACCTAATAGC ACCACATCAA -3'
IndCat-OQ2	27	$1\bar{3}t$ $\bar{1}2$ $\bar{1}1$ Q	5'- ATCGAGA TTGATGTGGT GCTATTAGGT /IAbRQ/ -3'

Table 4-S3: Independent input/output catalyst system sequences

Modularity is facilitated by sequence independence of the input (catalyst) and output (product); it is for this reason that we do not label strand SB also an output, even though it is also catalytically released by C. The design of the catalysis reaction enforces some degree of sequence similarity between strands SB and C, and this limits its usefulness in the construction of larger-scale circuits.

4-S11. Cross-catalytic feedback circuit

The main text referred to experiments on a cross-catalytic feedback circuit. A schematic of the circuit is shown in Fig. 4-S10A. Recall that domains $2a$, $4b$, and $9b$ are identical to each other, and renamed x here. Domains $2b$, 5, and 10 are identical to each other, and renamed y here. Thus, the design is actually very similar to that of the feedforward quadratic circuit, the only difference being that the the $OB1$ output strand has been replaced by $OB1E$, possessing the $9a$ domain at its 5'-most end, rather than the 1 domain.

Kinetic data from fluorescence studies is shown in Fig. 4-S10B. The cross-catalyst system is more leaky than the autocatalyst system, and 1% trigger is indistinguishable from the untriggered reaction. The reactions used in the modelling of the cross-catalyst system are shown below:

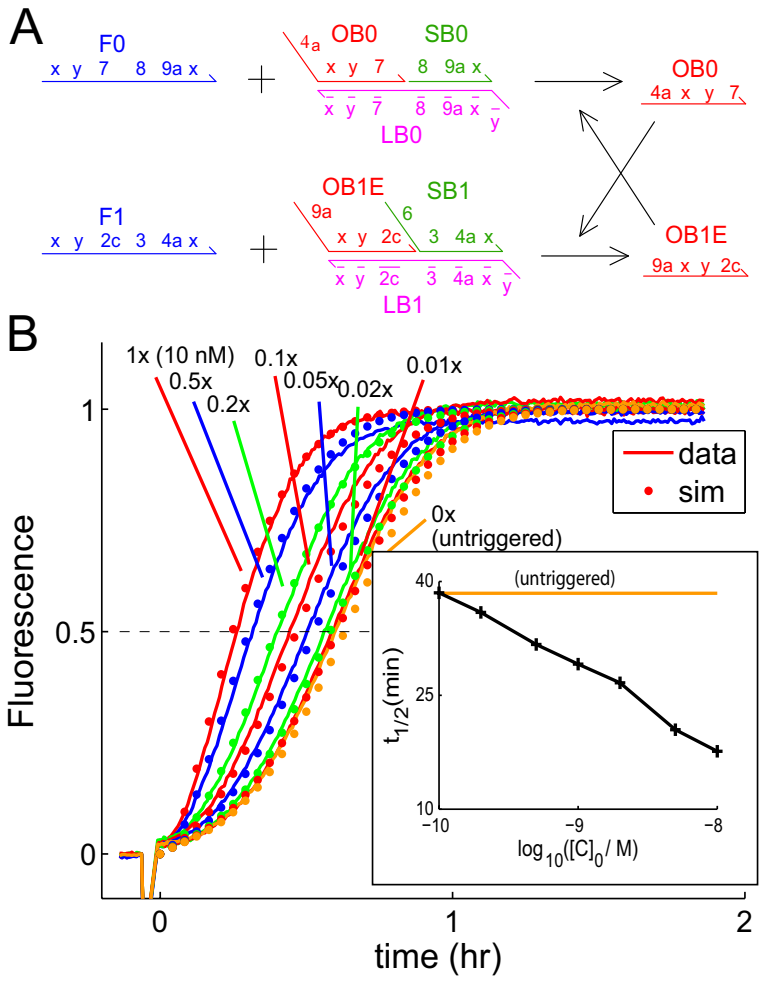


FIG. 4-S10: Cross-catalyst circuit **(A)** Cross-catalyst schematic. **(B)** Cross-catalyst series. Fluorescence studies were performed using the *SR* reporter complex shown in Fig. 4-S3A (same as that used for autocatalyst). Only catalyst *OB1E* was added at t $\approx$ 0. **(Inset)** Semi-log plot of time to reaction half completion as a function of catalyst added.

$$S0 + F0 \overset{k_{00}}{\rightarrow} OB0 + SB0 + W0 \tag{21}$$

$$S0 + OB1E \underset{k_{-10}}{\overset{k_{10}}{\rightleftharpoons}} I30 + SB0 \tag{22}$$

$$I30 + F0 \overset{k_{20}}{\rightarrow} I50 + OB0 \tag{23}$$

$$I50 \underset{k_{-30}}{\overset{k_{30}}{\rightleftharpoons}} OB1E + W0 \tag{24}$$

$$S1 + F1 \overset{k_{01}}{\rightarrow} OB1 + SB1 + W1 \tag{25}$$

$$S1 + OB0 \underset{k_{-11}}{\overset{k_{11}}{\rightleftharpoons}} I31 + SB1 \tag{26}$$

$$I31 + F1 \overset{k_{21}}{\rightarrow} I51 + OB1E \tag{27}$$

$$I51 \underset{k_{-31}}{\overset{k_{31}}{\rightleftharpoons}} C1 + W1 \tag{28}$$

$$\tag{29}$$

The only difference between the cross-catalytic circuit and the feedforward circuit is the identity of the 5' domain of the $OB1$ strand; this difference should not affect any rate constants, so the rate constants used in simulations are exactly the same as those used for the quadratic feedforward circuit. The only additional parameter we need to fit is the effective initial concentrations of $OB0$ and $OB1E$. Fig. 4-S10B shows the fits with initial concentrations $[OB0] = [OB1E] = 280$ pM. It is not clear why the impurity of these substrates is higher than that of the quadratic feedforward circuit.

4-S12. Entropy-driven catalytic analog AND gate

In this section, we demonstrate a reaction mechanism for a catalytic analog AND gate. Fig. 4-S11A shows a reduced schematic of the function of the AND gate: Fuel strands 32-33-34-1 and 2-3-4 (not shown) displace the 5' and 3' regions of the output strand 34-1-2 from the respective linking strands, and these reactions are catalyzed by strands 31-32 and 4-5. When both the 5' and 3' ends of the output strand are released, it reacts with the reporter complex (same as in Fig. 4-1D of main text) and fluorescence increases.

The multiplicative (AND-like) behavior can be understood quantitatively as follows.

The left (5') and right (3') catalytic ends operate independently and follow approximately the same kinetics as the catalyst system of Fig. 4-1. Let $f_L(t)$ be the fraction of left ends that have reacted with the fuel strand 32-33-34-1, and let $f_R(t)$ be the fraction of right ends that have reacted with the fuel strand 2-3-4. Then the fraction of output strand that has been released on both ends, and thus made active, is $f_{out}(t) = f_L(t) \cdot f_R(t)$. At initial times, when catalytic activity is linear in catalyst strand concentrations x and y, we thus have $f_{out}(t) \approx Mxyt^2$ for some constant M. Consequently, at a fixed time (prior to saturation), the output concentration is proportional to the product of the input concentrations.

Dom.	Sequence	Length
31	5'- CACACA -3'	6
32	5'- ACTTCAGTCATTAAGC -3'	16
33	5'- AGAC -3'	4
34	5'- CCATACAAGTATCA -3'	14

Table 4-S4: New domain sequences for the catalytic analog AND gate

As the 3' region of the output and the substrate are very similar to the system given in Fig. 4-1, the same reaction rates were used here for simulation. For the reaction rates relevant to the 5' catalytic component, the same rate constants were used as for the quadratic feedforward circuit, taken as "typical" values. That is, the reactions rates were not fitted to the data given here, even though the reaction rates most likely differ, since the sequences for the left end of the AND gate are quite different from those of the upstream catalyst of the quadratic circuit. New domain sequences are given in Table 4-S4.

[1] R. F. Gesteland, T.R. Cech, J.F. Atkins, *The RNA World* (2006).

[2] N. C. Seeman, *Trends Biochem. Sci.* **30**, 19 (2005).

[3] J. Bath, A. J. Turberfield, *Nature Nanotech.* **2**, 275 (2007).

[4] G. F. Joyce, *Annual Review of Biochemistry* **73**, 791 (2004).

[5] A. J. Turberfield, J. C. Mitchell, B. Yurke, A. P. Mills, M. I. Blakey, F. C. Simmel, *Phys. Rev. Lett.* **90**, 118102 (2003).

[6] J. S. Bois, S. Venkataraman, H. M. T. Choi, A. J. Spakowitz, Z. G. Wang, N. A. Pierce, *Nucleic Acids Research* **33**, 4090 (2005).

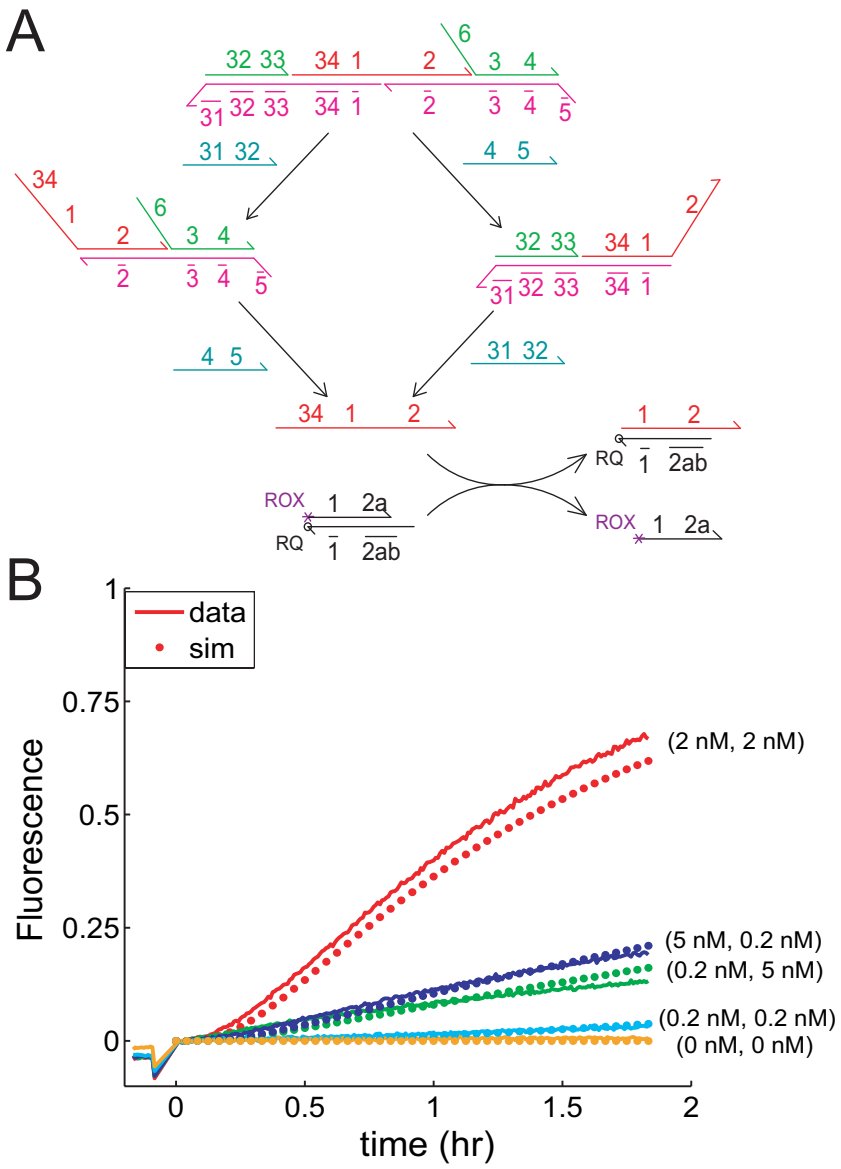


FIG. 4-S11: Catalytic AND gate. **(A)** Schematic of function. The output strand (34-1-2) is sequestered on both the 5' and 3' ends in the substrate. The catalysts 31-32 and 4-5 function independently to release the 5' and 3' ends, respectively, of the output. This action requires fuel strands 32-33-34-1 and 2-3-4 (not pictured). Only when both ends are released is the output strand able to react with the reporter complex. (In the right-hand pathway, single-stranded domain 2 can interact with the reporter complex, but initiation of four-way branch migration through helical domains 1, which could in principle complete triggering of the reporter, is sufficiently slow as to be negligible in practice.) The design is symmetric, despite appearances; domains 34 and 1 always appear together, and their lengths sum to the same as that of domain 2 (24 nt). (They labeled separately only for historical reasons to clarify interactions with the reporter complex.) **(B)** Fluorescence verification of catalytic AND gate function. $[S] = 10$ nM, $[F1] = [F2] = 13$nM. ($[C1]$, $[C2]$) = (2 nM, 2 nM) is more effective at releasing output than (5 nM, 0.2 nM) and (0.2 nM, 5 nM) even though the latter combinations possess higher total catalyst concentration.

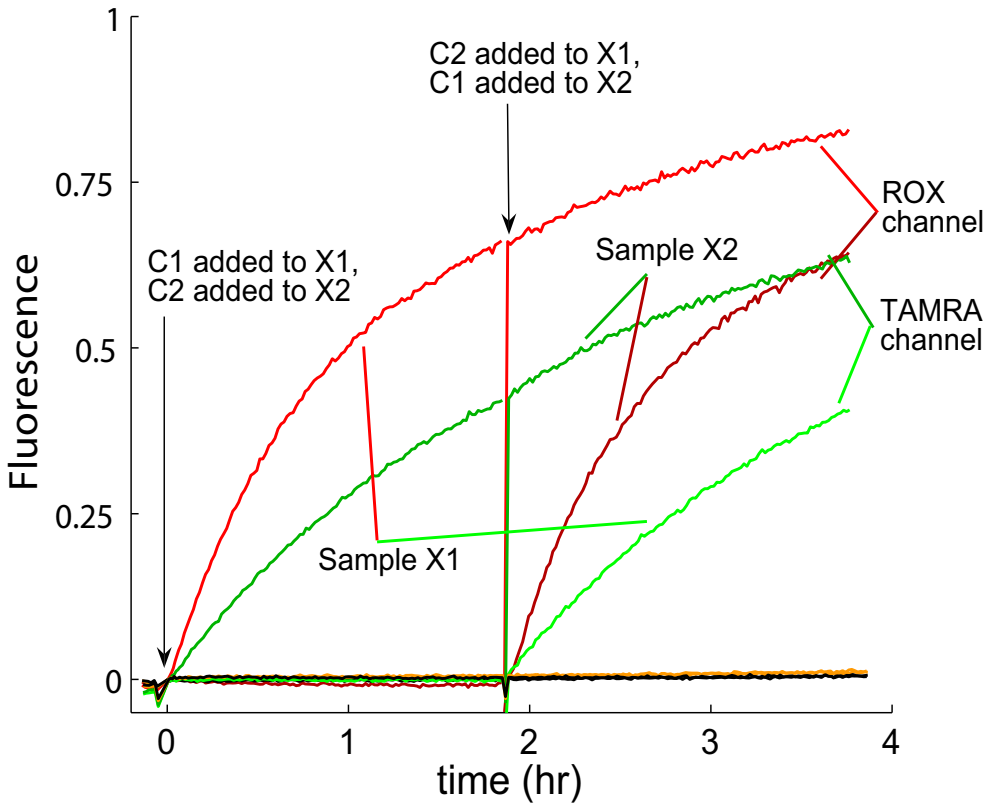


FIG. 4-S12: Simultaneous function of two unrelated catalytic systems. The red and green traces show the same sample ($X1$), monitored in the ROX and TAMRA channel, respectively. The dark red and dark green traces show the same sample ($X2$), monitored in the ROX and TAMRA channel, respectively. Fuels for both the system presented in Fig. 4-1 ($F1$) and in Fig. 4-S9 ($F2$) were present in both samples from the beginning. At $t \approx 0$, 10 nM (1x) $S1$ and $S2$ were added to both $X1$ and $X2$. Additionally, at $t \approx 0$, 0.1x $C1$ was added to $X1$, while 0.1x $C2$ was added to $X2$. Accordingly, the red and dark green traces showed increase in fluorescence due to catalytic activity, while the dark red and green traces show that catalysts $C1$ and $C2$ do not possess unwanted catalytic behavior (by catalyzing the other reaction). Control experiments showed that the ROX fluorophore in isolation is detected on the TAMRA channel with brightness 0.1112 relative to signal as detected on the ROX channel. Similarly, the TAMRA fluorophore in isolation produces signal on the ROX channel with efficiency 0.0687 relative to the signal detected on the TAMRA channel. The traces shown in this figure have been adjusted to remove fluorophore channel bleeding (using the data points between t = 0 and 1.8 hr on the red and dark green traces as references). At $t \approx 1.8$ hr, 0.1x $C1$ was added to $X2$, and 0.1x $C2$ was added to $X1$. The green and dark red traces then show increased fluorescence activity, showing that presence of other catalyst systems does not inhibit the proper function of catalysis.

[7] G. Seelig, B. Yurke, E. Winfree, *J. Am. Chem. Soc.* **128**, 12211 (2006).

[8] S. J. Green, D. Lubrich, A. J. Turberfield, *Biophysical Journal* **91**, 2966 (2006).

[9] M. N. Stojanovic, T. E. Mitchell, D. Stefanovic, *J. Am. Chem. Soc.* **126**, 3555 (2002).

[10] J. Macdonald, Y. Li, M. Sutovic, H. Lederman, K. Pendri, W. Lu, B. L. Andrews, D. Stefanovic, M. N. Stojanovic, *Nano Letters* **6**, 2598 (2006).

[11] H. Lederman, J. Macdonald, D. Stefanovic, M. N. Stojanovic, *Biochem.* **45**, 1194 (2006).

[12] M. Hagiya, S. Yaegashi, K. Takahashi, *Nanotechnology: Science and Computation*, 293-308 (2006).

[13] M. Levy, A. D. Ellington, *Proc. Nat. Acad. Sci.* **100**, 6416 (2003).

[14] R. M. Dirks, N. A. Pierce, *Proc. Nat. Acad. Sci.* **101**, 15275 (2004).

[15] M. N. Stojanovic, S. Semova, D. Kolpashchikov, J. Macdonald, C. Morgan, D. Stefanovic, *J. Am. Chem. Soc.* **127**, 6914 (2005).

[16] R. Penchovsky, R. R. Breaker, *Nature Biotechnology* **23**, 1424 (2005).

[17] G. Seelig, D. Soloveichik, D. Y. Zhang, E. Winfree, *Science* **314**, 1585 (2006).

[18] In the system presented in Fig. 4-1, there is some sequence redundancy in the domain sequences chosen (for example, 2*b* and 5 are identical). This is because all four systems presented in this paper were designed together, with the goal of minimizing the number of differences between systems. In designing a catalytic reaction in isolation, there are no sequence constraints; we demonstrate a system with completely independent catalyst and output in text 3-S10 [22]; it possesses very similar kinetics to the reaction shown in Fig. 4-1.

[19] M. Zuker, *Nucleic Acids Res.* **31**, 3406 (2003).

[20] J. Sager, D. Stefanovic, *DNA Computing 11*, LNCS **3892**, 243 (2006).

[21] K. U. Mir, *DNA Based Computers II*, DIMACS **44**, 275 (1999).

[22] See supporting material on *Science* Online.

[23] B. Yurke, A. P. Mills, *Genetic Programming and Evolvable Machines* **4**, 111 (2003).

[24] I. G. Panyutin, P. Hsieh, *J. Mol. Bio* **230**, 413 (1993).

[25] S. A. Marras, F. R. Kramer, S. Tyagi, *Nucleic Acids Research* **30**, 122 (2002).

[26] C. Green, C. Tibbetts, *Nucleic Acids Research* **9**, 1905 (1981).

[27] R. Higuchi, C. Fockler, G. Dollinger, R. Watson, *Nature Biotech.* **11**, 1026 (1993).

[28] N. Paul, G. F. Joyce, *Current Opinion in Chemical Biology* **8**, 634 (2004).

[29] Yin, P., Choi, H. M. T., Calvert, C. R. & Pierce, N. A. Programming biomolecular self-assembly pathways. *Nature* **451,** 318-322 (2008).

[30] Y. Benenson, B. Gil, U. Ben-Dor, R. Adar, E. Shapiro, *Nature* **429**, 423 (2004).

[31] R. Pei, S. K. Taylor, D. Stefanovic, S. Rudchenko, T. E. Mitchell, M. N. Stojanovic, *J. Am. Chem. Soc.* **128**, 12693 (2006).

[32] X. Li, D. R. Liu, *Angew. Chem. Int. Ed.* **43**, 4848 (2004).

[33] A. D. Ellington, J. Szostak, *Nature* **346**, 818 (1990).

[34] J. Tang, R. R. Breaker, *Chemistry & Biology* **4**, 453 (1997).

[35] S. Muller, D. Strohbach, J. Wolf, *Nanobiotechnology* **153** 31 (2006).

[36] F. J. Isaacs, D. J. Dwyer, J. J. Collins, *Nature Biotechnology* **24** 545 (2006).

[37] K. U. Mir, *Proc. DNA Based Computers II* **44**, 243 (1999).

[38] M. Zuker, *Nucleic Acids Res.* **31**, 3406 (2003).

[39] J. D. Puglisi, I. Tinoco, *Methods in Enzymology* **180**, 304 (1989).

[40] M. Torimura, S. Kurata, K. Yamada, T. Yokomaku, Y. Kamagata, T. Kanagawa, R. Kurane, *Anal.*

Sci. **17**, 155 (2001).

[41] J. SantaLucia, *Proc Natl Acad Sci USA* **95**, 1460 (1998).

Chapter 5: Robustness and Specificity of the DNA catalyst

Imagine that every time you turn on your computer, your microwave and dishwasher also start. When you try to turn off the microwave, the air conditioner and the alarm clock also lose power. When you try to restart the air conditioner, the stoves and the blender also fire up. Life would be pretty annoying, and you may even consider throwing out all your electical equipment to live the life of a Luddite.

Fortunately, electrical appliances and tools do not typically interfere with each other in the way I just described. Each appliance functions *modularly,* not affecting or being affected by the function of others. Electrical appliances are easy to make modular because the appliances are usually separated by space, and share only a power line (and even just that sometimes causes problems... e.g. you might notice your lights dim if you turn on a really powerful vacuum cleaner). In the aqueous nanoscale systems we are examining, molecules necessarily bump into other molecules continually, and the potential for undesireable side reactions increase the difficulty of designing modular reactions and devices.

Here, we examine the modularity and robustness properties of the DNA catalyst system presented in the previous chapter, with the primary goal being to discovering how well the mechanism would work in an environment rife with other biomolecules. The other goal of this work was to explore the importance of DNA purity, as a practical economic

consideration for using the catalytic amplification mechanism.

I came up with the initial idea for this work in 2007 while preparing the manuscript describing the catalyst shown in Chapter 3. The sheer number of experiments needed to fully characterize the DNA catalyst system was quite daunting initially, but I finally managed to finish in October 2009. Then, with record speed, Erik and I managed to agree on the manuscript wording and submit it for publication November of the same year. Erik Winfree helped model the system, analyze the data, and revise the manuscript. This work was submitted for publication on November 18, 2009 as:

Zhang, David Yu; Winfree, Erik. "Robustness and Modularity Properties of a Non-covalent DNA Catalytic Reaction."

The biophysics of nucleic acid hybridization and strand displacement have been used for the rational design of a number of nanoscale structures and functions. Recently, molecular amplification methods have been developed in the form of non-covalent DNA catalytic reactions, in which single-stranded DNA molecules catalyze the release of single-stranded DNA product molecules from multi-stranded complexes. Here, we characterize the robustness and specificity of one such strand displacement-based catalytic reaction. We show that the designed reaction is simultaneously sensitive to sequence mutations in the catalyst and robust to a variety of impurities and molecular noise. These properties facilitate the incorporation of strand displacement-based DNA components in synthetic chemical and biological reaction networks.

Nucleic acids have been identified and demonstrated as versatile nanoscale engineering materials because of their specific binding properties and their well understood thermodynamics [1] and kinetics [2, 3]. Recent constructions exhibiting dynamic behavior include logic gates and networks [4–6], motors [7, 8], and amplification mechanisms [9–15]. By perfecting the design of these and other modular primitives, nucleic acid engineering may one day allow precise spatial-temporal control of chemistry [16–19]. Such "molecular program-

ming," if made sufficiently robust, could be incorporated into biological cells and organisms to allow the dynamic programming of development and behavior.

In order for such dreams to be realized, however, the basic nucleic acid components being developed must approach a level of modularity and robustness comparable to those of transistors and other analogs from electrical engineering. Robustness denotes a low amount of interference between the device and the environment. For aqueous molecular computation applications, robust nucleic acid constructions must (1) function over a broad range of ambient solution conditions, (2) function despite imperfections and impurities, and (3) function in the presence of external molecular noise (in the form of other biomolecules that may exist in solution). Modularity denotes a low amount of interference between devices. For nucleic acid devices, modularity requires a generalized, algorithmic method of component construction such that different components interfere with each other minimally. In this work, we explore and evaluate the robustness and modularity properties of one particular DNA component, so that we can appropriately incoporate it in large and/or noisy systems.

One key component needed for generalized molecular computation is the signal amplifier. In biology as well as in electrical engineering, signal amplification serves a key role in the robustness of networks and kinetics of signal transduction. Nucleic acid signal amplifiers have been implemented on a molecular level with the construction of non-covalent catalytic reactions, in which single-stranded molecules of DNA (henceforth *strands*) catalytically release product strands from multi-stranded complexes [9, 13–15]. To concisely describe the functions of and reactions involving each strand, we subdivide strands into *domains*, contiguous bases that act as an abstract unit in binding and dissociation (Fig. 5-1A) [9].

Fig. 5-1B shows the designed pathway of the catalytic reaction presented in ref. [9] using domain notation. In the illustration, strand C binds to multi-stranded substrate complex S to form intermediate $I1$, which subsequently reacts with fuel strand F to release output product OP, side product SP, and the catalyst C. In the absence of C, the reaction between F and S is very slow. Thus, input strand C dynamically amplifies output OP; this is roughly analogous to a NPN transistor where F and S together act as the collector, C

as the base, and OP as the emitter.

Crucial to the function of this catalytic reaction are toeholds, short single-stranded DNA domains that initiate binding between DNA strands and complexes. In Fig. 5-1B, the reaction between catalyst C and substrate S is initiated by the hybridization of toehold domain 5 on C to its complement $\bar{5}$ on S. The binding energy of toehold domains have been shown to sensitively affect the kinetics of strand displacement reactions [2, 3].

The catalyst C can potentially be any single-stranded oligonucleotide with limited secondary structure: the sequences of substrate S and fuel F are designed so that their reaction is catalyzed by strand C of a given sequence. Strand C is abstracted here as the concatenation of the two domains 4 and 5, such that the toehold domain 5 is the six 3'-most nucleotides, and domain 4 contains the remaining bases in C. The kinetics of this catalyst system are fastest when C and F possesses no secondary structure; significant secondary structure in C or F will slow the kinetics of the reaction because C and/or F must spontaneously unfold before they can react with S and $I1$, respectively.

Previous work established that the reaction shown in Fig. 5-1B is robust to the temperature and salt concentration changes [9]. Here, we characterize other robustness and modularity properties of this molecular amplification mechanism, including:

Maximum catalytic turnover. The maximum turnover is the average number of reactions catalyzed by each catalyst molecule, and determines the length of time that the system proceeds as designed. High turnover is necessary for sustained robust function.

Effects of overhangs on the catalyst molecule. The catalyst molecule in our system is approximately 20 nt long. For amplifying and modularly interfacing with longer nucleic acids, unique subsequences could be used if overhangs do not significantly affect catalytic activity.

Importance of the 5'/3' orientation. Proper function of the mechanism with inverted 5'/3' orientation allows greater flexibility in interfacing with other nucleic acid systems, enhancing modularity.

Sensitivity to mutations in the fuel and catalyst. High reaction specificity would ensure that only the desired molecule triggers the programmed reaction, and allows for the

simultaneous operation of many different modular subsystems.

Effects of strand impurities. Imperfect synthesis of oligonucleotides can lead to reactions other than those of the desired pathway. While impurities can be removed via purification, systems robust to strand impurities have a clear advantage.

Interference from DNA molecules of randomized sequence. Sensitivity to crosstalk from unrelated molecules would reduce modularity in engineered systems and undermine robust function in environments containing biological molecules.

Sequence robustness of the design. Stringent sequence design requirements would limit the number of DNA components that can simultaneously be in solution, and reduce the modularity and robustness of the design.

In essence, we are interested in characterizing the properties of this particular mechanism so that it can be used as a "plug and play" module for incorporation into generalized chemical reaction networks. While the properties of other DNA strand displacement-based devices vary somewhat by design, it is hoped that many of these components possess similar robustness and modularity properties, so that the results presented in this work are generalizable.

Materials and Methods

Buffer conditions. The buffer for all experiments was TE (10 mM Tris · HCl pH balanced to 8.0, 1 mM EDTA), purchased as 100x stock (Sigma-Aldrich), with 12.5 mM $MgCl_2$ added. Because EDTA chelates magnesium ions, the effective concentration of Mg^{2+} is 11.5 mM. All experiments and purifications were performed at 25 °C.

DNA sequences and design. The sequences of the basic catalyst system shown in Fig. 5-1B are the same as those in [9], which were carefully designed by hand to avoid secondary structure in single-stranded species (Table 5-1). To this end, one useful sequence design heuristic was to minimize the number of G's in single-stranded species (e.g. C, F, and OP). Not only are unintentional G-C bases much more stable than A-T ones, G's also further contribute to undesired secondary structures through G-T wobbles and G-quartets [25]. For this reason, there are a total of only 4 G's in domains 1 through 6

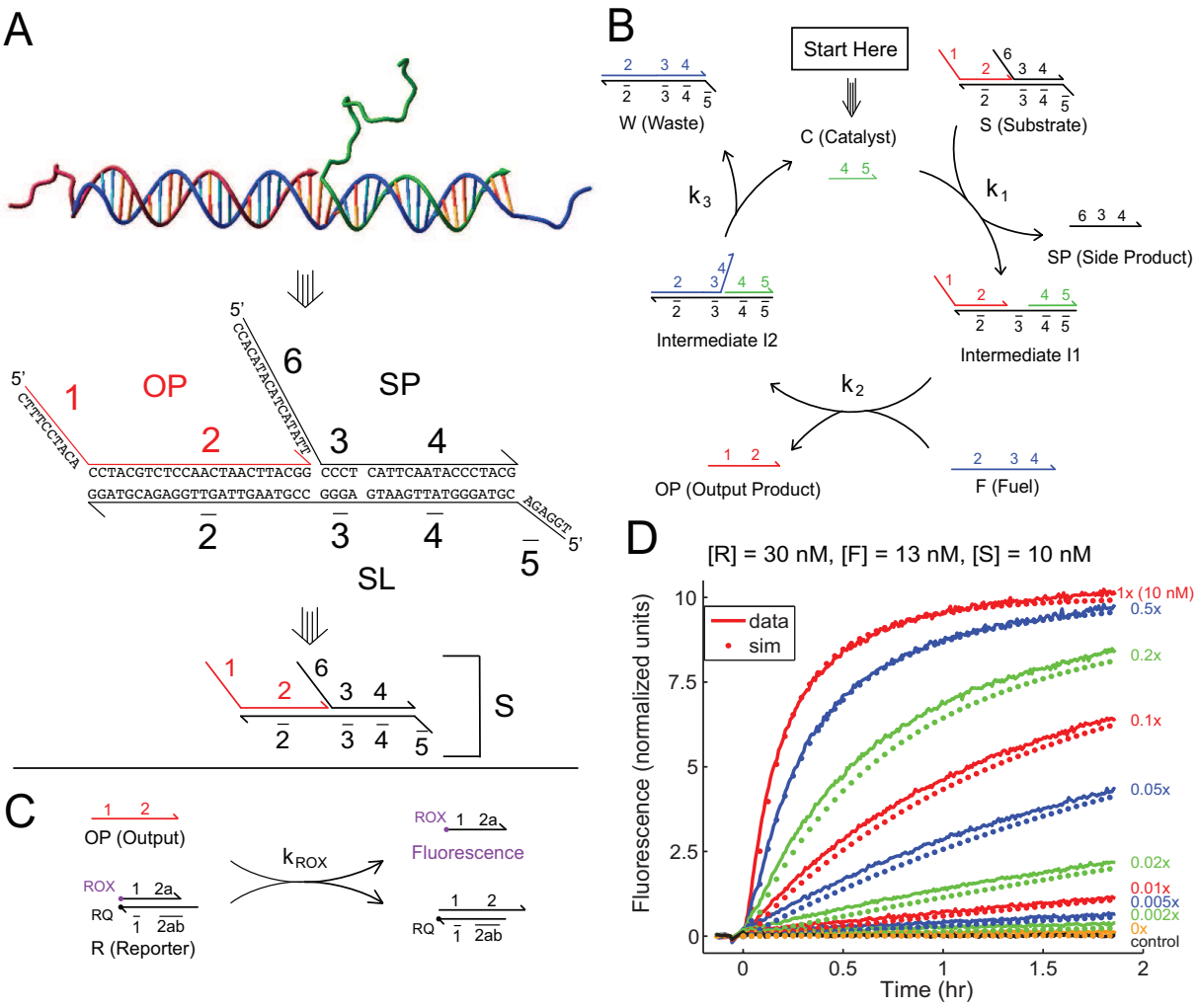


FIG. 5-1: A non-covalent strand displacement reaction catalyzed a target single-stranded DNA molecule C (adapted from ref. [9]). **(A)** DNA abstraction. The double-helix DNA molecule (top) is typically abstracted as two directional lines, one for each strand, with base identities shown (middle). Here, we abstract the DNA molecule one step further by grouping contiguous nucleotides into domains, functional regions of DNA that act as a unit in binding (bottom). Domains are labelled by numbers. Domain $\bar{x}$ is the complement of (and will hybridize to) domain x. The strands OP, SP, and SL form the 3-stranded DNA complex S. The DNA molecule in the top panel was drawn using Nanoengineer, a free DNA visualization software by Nanorex. **(B)** The designed mechanism of catalytic function. **(C)** Fluorescent reporter complex. Output product OP reacts stoichiometrically with reporter complex R to yield a fluorescent strand. ROX denotes the carboxy-X-rhodamine fluorophore (attached to the DNA molecule via an NHS ester), and RQ denotes the Iowa Black Red Quencher. This indirect reporter complex was used because of the thermodynamic effects of fluorophore-quencher binding [20]. From ref. [9], k_{ROX}, the second-order rate constant of reaction between OP and R, was measured to be $4 \cdot 10^5$ M^{-1} s^{-1}. In experiments, the concentration of the reporter R was in excess of the concentration of the fuel F and substrate S to minimize the reporter delay (no more than two minutes for $[R] = 30$ nM). **(D)** Experimental and simulation results from ref. [9]. Dotted lines show ordinary differential equation (ODE) simulation results according to the model in Table 5-2.

(out of 76 bases). The sequences for the 5'/3'-inverted catalyst were similarly designed to minimize G content. Domains 10 through 13 possess only 3 G's (out of 40 bases).

The sequences used for the 4-letter alphabet catalyst system shown in Figs. 10 and 11 were designed by hand to possess minimal secondary structure while balancing the base composition of the domains (Table 5-5): Domains 41 through 46 possess 16 G's, 16 C's, 26 A's, and 21 T's. Domains 51 and 52 possess 7 G's, 10 C's, 14 A's, and 6 T's. Domains 63 through 65 were also designed to minimize the number of G's, and together possess 1 G, 11 C's, 7 A's, and 5 T's. Domains 71 through 76, being the 5'/3' inverse of domains 1 through 6, have the same nucleotide distribution as the latter. The theoretical and experimental effectiveness of using three-letter and four-letter alphabets are discussed in the text.

Substrate purification. DNA oligonucleotides used in this study were purchased from Integrated DNA Technologies (IDT), purified by high performance liquid chromotography (HPLC), except as noted in the text. Where applicable, fluorophores and quenchers were attached by IDT as well.

We further purified all multi-stranded complexes (i.e. substrates S and the reporter complexes R) by non-denaturing (ND) polyacrylamide gel electrophoresis (PAGE) as follows: Strands for each sample were prepared at 20 μM concentration, subject to pipetting, dilution, and extinction coefficient errors. The samples were then annealed with an Eppendorf Mastercycler Gradient thermocycler, cooling from 95 °C to 20 °C at a constant rate over the course of 90 minutes.

ND loading dye containing Xylene Cyanol FF (XCFF) in 50% glycerol was added to all samples, achieving final gycerol concentration of 10% by volume. The samples were run on 12% ND PAGE at 180 V for 6 hours at 25 °C (using a Novex chamber with external temperature bath).

The proper bands were cut out and eluted in 1 mL TE/Mg^{2+} buffer for 2 days. Typical yields ranged from 30% to 60%. Purified complexes were quantitated by measurement of absorbance at 260 nm using an Eppendorf Biophotometer, and concentrations were calculated using extinction coefficients for single- and double-stranded DNA predicted by nearest-neighbor models [22].

Denaturing PAGE. A 12% acrylamide (19:1 acrylamide:bis) denaturing PAGE was run to prepare Fig. 5-8C. Acrylamide solution was diluted from 40% stock acrylamide (Ambion).

Denaturing loading dye containing XCFF in 80% formamide was added to all samples in 1:1 ratio, achieving final formamide concentration of 40% by volume. Gels were run at 120 V for 1 hour at 25 °C, with temperature controlled using a Novex chamber external temperature bath. Gels were stained with Sybr-Gold stain (Invitrogen), and scanned with a Bio-Rad Molecular Imager.

Spectrofluorimetry studies. Spectrofluorimetry studies were done using a SPEX Fluorolog-3 (Horiba) with 1.6 mL synthetic quartz cells (Hellma catalog number 119-004F). Sample solutions were excited at 588 nm, and emission at 602 nm was observed (optimal signal for ROX fluorophore in our buffer). Slit size used were 2 nm for both excitation and emission monochromators. Fluorescence experiments were done with integration time of 10 seconds for every 60 second time-point.

Prior to each experiment, all cuvettes were cleaned thoroughly: each cuvette was washed 15 times in distilled water, once in 70% ethanol, another 5 times in distilled water, and finally once more in 70% ethanol. Appropriate volumes of DNA stock solutions, typically 1 μM in concentration, were added to TE/Mg^{2+} buffer to achieve the correct final concentrations with a total volume of 1.5 mL in the cuvettes. Solutions were mixed by pipetting 250 μL of solution in and out of the cuvette 32 times. Stir bars were not used because they have been observed to significantly contribute to fluorophore decay. Unpublished data indicate that this may be due to bleach (used for cleaning) slowly desorbing from the surface of the stir bars, that in turn cause nonspecific fluorophore quenching [24].

For the slit size, concentrations, and times chosen, no measurable photobleaching was observed. Fluorescence measurements are linear in the concentration of the free fluorescent strand F. All experimental results were within the linear regime of the spectrofluorimeter detector, according to specification sheets provided by the manufacturer.

Fluorescence normalization. Fluorescence is normalized so that 1 normalized unit (n.u.) of fluorescence corresponds to 1 nM of unquenched fluorophore-labelled strand F. This normalization is based on the fluorescence levels of annealed samples: a negative control with only $[R] = 30$ nM (normalized to 0 n.u.), and a positive control with $[R] = 30$ nM, $[F] = 20$ nM and $[S] = 10$ nM (normalized to 10 n.u.). Normalization experiments were

run once for each different purified reporter complex. Day-to-day and sample-to-sample variations are estimated to be less than 5% [9].

Carrier strands. It has been observed that DNA sticks non-specifically to pipette tips, so that serial dilutions lead to stocks more dilute than expected [9]. Unfortunately, this loss is not consistent, so we could not compensate for tip loss with additional reagent. Instead, we introduced into all dilute stocks (1 μM and below) a non-reactive 20 nt poly-T "carrier" strand, at a concentration of 1 μM. Because pipette tip loss is non-specific, the majority of DNA loss would be of the carrier strand, and serially diluted stocks are only slightly more dilute than expected. Poly-T strands have minimal influence on the reactions of other DNA molecules in this system [9].

Parameter Fitting. The best-fit rate constants to the revised ODE model of the system (Table 5-3) were fitted using the 'fminunc' function in Matlab to minimize the error between experimental data and the reaction model. The error is calculated as follows:

$$\text{Error} = \sum_{t,\text{traces}} \frac{(F_d(t) - F_m(t))^2}{t_{max} \cdot [S]_0^2}$$

where $F_d(t)$ is the fluorescence value of the data at time t, and $F_m(t)$ is the fluorescence value predicted by the ODE model at time t. The denominator terms correspond to a normalization factor to ensure that each plot contributes roughly equally in fits: t_{max} denotes the endpoint time of each experiment, and $[S]_0$ denotes the initial concentration of the substrate S (the maximum fluorescence level of a catalytic reaction that achieves 100% completion). The rate constants' confidence intervals were determined as the values at which the error value would double, holding all other parameters constant.

Results and Discussion

Turnover Characterization and Revised Modeling

The *turnover* is the number of reactions catalyzed by each catalyst molecule over a particular period of time. We can calculate the turnover from our experimental data by dividing the excess number of product molecules (over the uncatalyzed reaction) by the

Dom.	Sequence	Length (nt)
1	5'- CTTTCCTACA -3'	10
2a	5'- CCTACG -3'	6
2b	5'- TCTCCA -3'	6
2c	5'- ACTAACTTACGG -3'	12
2	5'- CCTACGTCTCCAACTAACTTACGG -3'	24
3	5'- CCCT -3'	4
4	5'- CATTCAATACCCTACG -3'	16
5	5'- TCTCCA -3'	6
6	5'- CCACATACATCATATT -3'	16
7	5'- TTCACCTCAGTTATG -3'	15
8	5'- TCAATTCCTAACATA -3'	15
9	5'- TTTTTTTTTTTTTTA -3'	15
10	5'- CACACA -3'	6
11	5'- ACTTCAGTCATTAAGC -3'	16
12	5'- AGAC -3'	4
13	5'- CCATACAAGTATCA -3'	14

Table 5-1: Domain sequences.
Sequences for domains 1 through 6 are from ref. [9]

number of catalyst molecules present [30]. In a perfect catalyst system, the turnover indefinitely increases linearly in time. In reality, however, the turnover will asymptotically approach a maximum value as time goes to infinity, because the catalyst will be inactivated by side reactions. In this section, we seek to characterize the maximum turnover of the DNA catalyst system, and to explore potential causes for this limitation.

Fig. 5-2A shows the catalytic activity using a high ratio of substrate to catalyst ($\frac{[S]}{[C]}$ between 200 and 1000), and Fig. 5-2B shows the inferred turnover. High substrate to catalyst ratios were used for this experiment so that turnover isn't limited by substrate quantity. The catalytic turnover for all three catalyst concentrations approached a value between 80 and 100 after 24 hours of reaction. Since the catalytic activity showed a consistent decreasing trend across all three experiments, it is likely that a real, but as yet undocumented, side reaction is inhibiting catalysis. The turnover after 24 hours is significantly lower than 200, 500, and 1000 (the turnover at 100% reaction completion for 0.005x, 0.002x, and 0.001x catalyst), so substrate depletion is unlikely to be the main cause of this slowdown.

One likely cause of limited turnover is DNA strand impurities. For example, a single deletion near the 3' end of the fuel strand F could slow the release of catalyst C from intermediate $I2$. With two or more deletions near the 3' end of the fuel F, the release of C from $I2$ could be thermodynamically unfavorable. Studies on the heterogeneity of

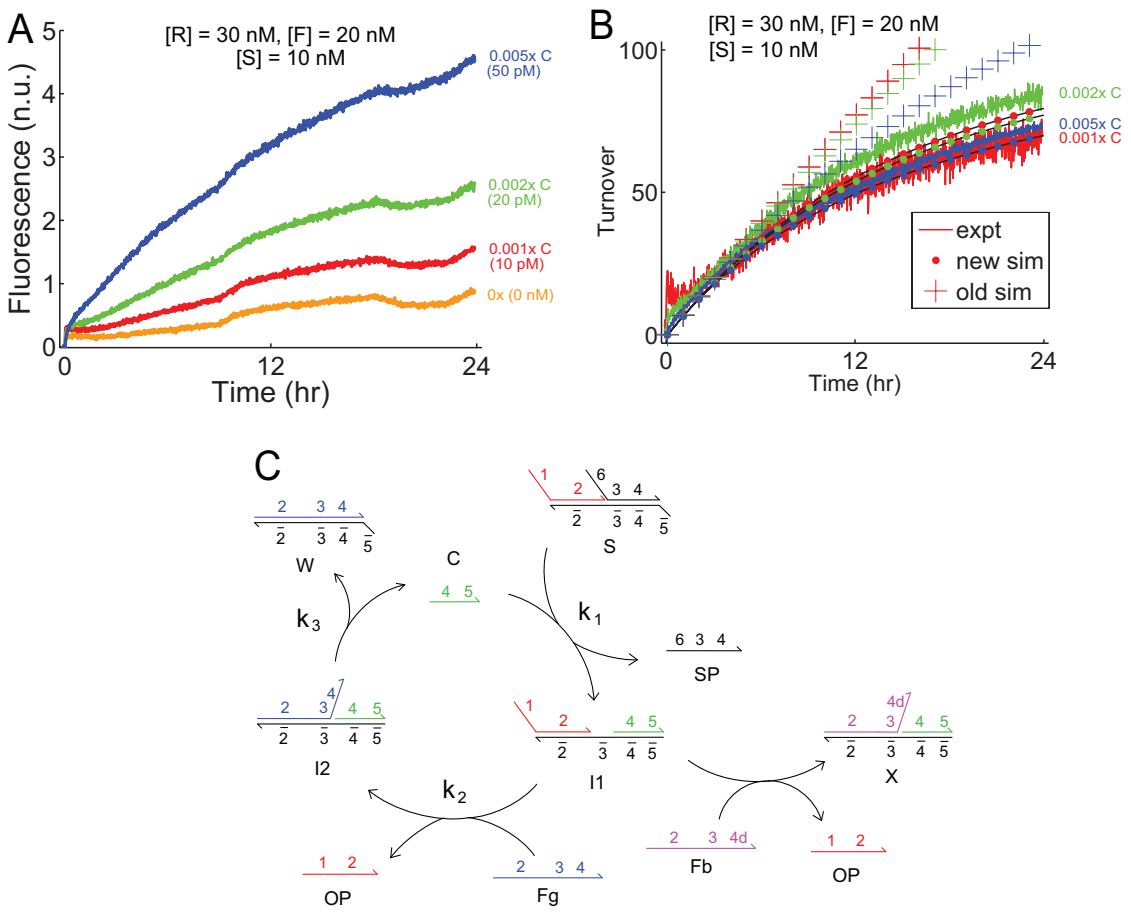


FIG. 5-2: Catalytic turnover. **(A)** Raw data for turnover experiments. Traces showed significantly more noise than typical; possibly, this is due to lamp and temperature instability. **(B)** Turnover plotted as a function of time. Turnover is calculated as the excess normalized fluorescence above leak (0x trace) divided by concentration of catalyst; e.g. $T_{0.001x}(t) = \frac{F_{0.001x}(t) - F_{0x}(t)}{0.001 \cdot 10 \text{ nM}}$. "Old sim" denotes simulations using the model presented in ref. [9], which was fitted only to the 10 nM data shown in Fig. 5-1D. The "new sim" simulations use the model and rate constants fitted to the data presented in this paper. The relative ordering and the quantitative differences between the "new sim" simulations and the experimental results are not considered significant–all traces are considered to be within experimental error of one another. **(C)** Our new model that accounts for F sub-populations. A small fraction (fitted to be 1.0%) of F exists as Fb, with deletions in the 4 domain. These react with $I1$ to yield X, from which C cannot dissociate.

commercially synthesized oligonucleotides [26] indicate that DNA oligonucleotide samples show deletions at every single base.

To model fuel impurities and observe whether such impurities would produce simulation results consistent with experimental data, we modified our model to assume that the total F concentration is the sum of two distinct fuel molecules: a "good" fuel Fg that reacts with $I1$ and release C from $I2$ as designed, and a "bad" fuel Fb that reacts with $I1$ to form a new product X from which C cannot dissociate (Fig. 5-2C). The relative concentrations of

$\mathrm{S} + \mathrm{C} \underset{k_2}{\overset{k_1}{\rightleftharpoons}} \mathrm{I1} + \mathrm{SP}$	$k_1 = 6.5 \cdot 10^5\ \mathrm{M}^{-1}\ \mathrm{s}^{-1}$
$\mathrm{I1} + \mathrm{F} \overset{k_2}{\rightarrow} \mathrm{I2} + \mathrm{OP}$	$k_2 = 4.2 \cdot 10^5\ \mathrm{M}^{-1}\ \mathrm{s}^{-1}$
$\mathrm{I2} \underset{k_1}{\overset{k_3}{\rightleftharpoons}} \mathrm{W} + \mathrm{C}$	$k_3 = 4 \cdot 10^{-3}\ \mathrm{s}^{-1}$
$\mathrm{S} + \mathrm{F} \overset{k_0}{\rightarrow} \mathrm{OP} + \mathrm{SP} + \mathrm{W}$	$k_0 = 2.3 \cdot 10^1\ \mathrm{M}^{-1}\ \mathrm{s}^{-1}$

Table 5-2: Original model

$\mathrm{S} + \mathrm{C} \underset{k_2}{\overset{k_1}{\rightleftharpoons}} \mathrm{I1} + \mathrm{SP}$	$k_1 = 2.7 \pm 0.5 \cdot 10^5\ \mathrm{M}^{-1}\ \mathrm{s}^{-1}$
$\mathrm{I1} + \mathrm{Fg} \overset{k_2}{\rightarrow} \mathrm{I2} + \mathrm{OP}$	$k_2 = 1.1 \pm 0.7 \cdot 10^6\ \mathrm{M}^{-1}\ \mathrm{s}^{-1}$
$\mathrm{I2} \underset{k_1}{\overset{k_3}{\rightleftharpoons}} \mathrm{W} + \mathrm{C}$	$k_3 = 1.1 \pm 0.5 \cdot 10^{-2}\ \mathrm{s}^{-1}$
$\mathrm{S} + \mathrm{Fg} \overset{k_0}{\rightarrow} \mathrm{OP} + \mathrm{SP} + \mathrm{W}$	$k_0 = 5\ \mathrm{M}^{-1}\ \mathrm{s}^{-1}$
$\mathrm{I1} + \mathrm{Fb} \overset{k_2}{\rightarrow} \mathrm{X} + \mathrm{OP}$ $\mathrm{S} + \mathrm{Fb} \overset{k_0}{\rightarrow} \mathrm{OP} + \mathrm{SP} + \mathrm{W2}$ $\mathrm{C} + \mathrm{W2} \overset{k_1}{\rightarrow} \mathrm{X}$	$\chi = \frac{[Fb]}{[Fg]+[Fb]} = 1.0 \pm 0.3 \cdot 10^{-2}$

Table 5-3: Revised model

Fb and Fg are believed to depend on synthesis and purification of the F strand, and thus will likely differ from sample to sample of F.

Our new model fits 5 parameters: the rate constants k_0, k_1, k_2, and k_3, as well as $\chi = \frac{[Fb]}{[Fb]+[Fg]}$, the fraction of bad fuel molecules with deletions near the 3' end. Best fit values under the new model are shown in Table 5-3. For comparison, the model presented in ref. [9] is reproduced in Table 5-2. This new model shows a significantly better quality of fit to the data in Fig. 5-2, 5-3, and 5-8, and implies a maximum achievable turnover of $\frac{1}{\chi} \approx 100 \pm 20$ for our current stock of F.

Fig. 5-3 shows the behavior of the catalyst system with various fuel and substrate concentrations ranging from 1 nM to 100 nM. The catalytic behavior is qualitatively similar for all of the tested concentrations, and our new model quantitatively predicts the kinetics for all of these experiments. Thus, the catalyst system functions reliably and predictably over at least 2 orders of magnitude of concentrations.

Overhangs in the Catalyst Input

One long-term goal of strand displacement-based nucleic acid devices and circuits is to interface with biological systems, taking biological nucleic acid molecules as inputs and producing active biomolecules as outputs. When using long biological nucleic acids (such as most mRNAs) as inputs, it is often desirable to use only a unique unstructured sub-

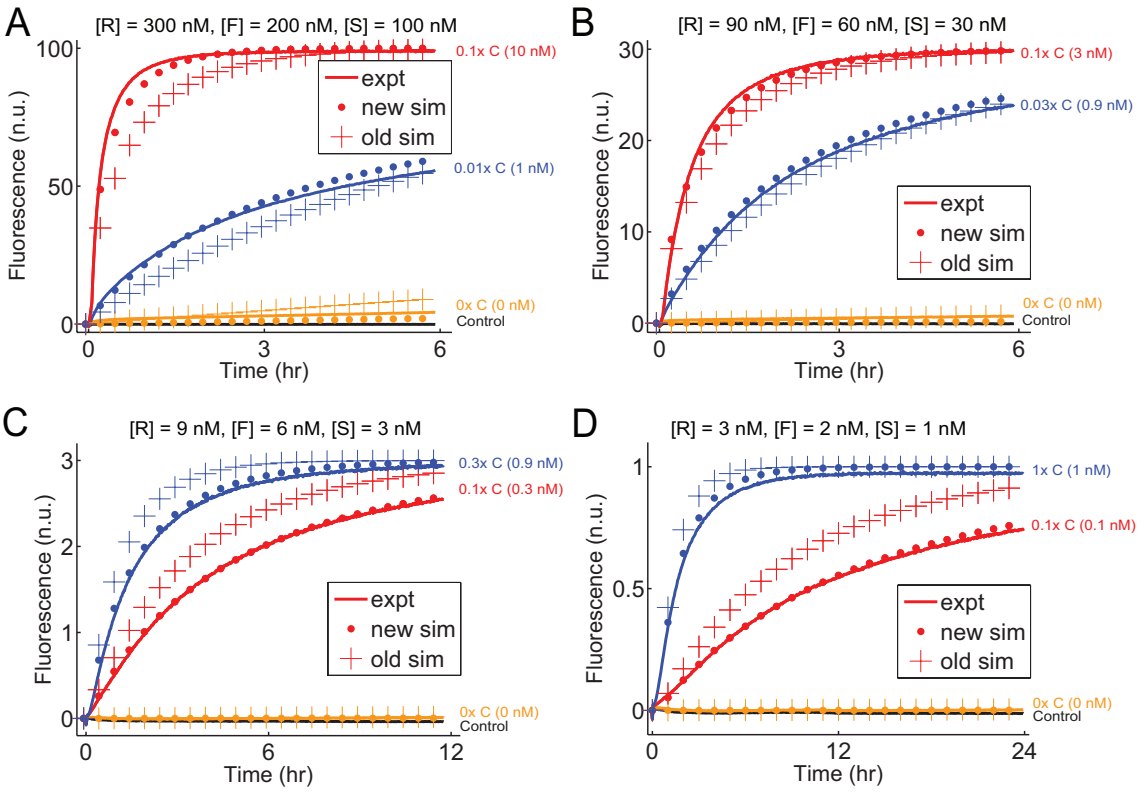


FIG. 5-3: Catalytic function and model results for **(A)** 100 nM, **(B)** 30 nM, **(C)** 10 nM, and **(D)** 1 nM substrate S concentration. The new model that accounts for catalyst inactivation (Table 5-3) fits the experimental data better than the old model (Table 5-2). Red traces denote 0.1x catalyst, while blue traces denote catalyst concentrations of about 1 nM.

sequence as the input to the synthetic DNA devices to improve kinetics, specificity, and cost. Consequently, the active input subsequence will usually possess 5'- and/or 3'- overhangs. Additionally, 5'- and 3'- overhangs may also be present in output signals of strand displacement-based DNA devices, as artifacts of upstream sources. For example, in ref [9], the output of the first catalytic reaction in the feed-forward network contained an extra 3' domain that served no catalytic purpose in the downstream reaction. Here, we study the effects of these overhangs on the kinetics of the studied catalytic reaction.

We constructed five alternative versions of the catalyst with single- and double-stranded 5'- and 3'- overhangs (Fig. 5-4ABCDE). The sequences of the 7 and 8 domains were designed to be minimally interactive with other domains existing in the reaction, while the 9 domain is mostly poly-T, and should exhibit the least interaction with other single-stranded domains (Table 5-1). Fig. 5-4 shows the catalytic activities of these five modified catalysts. 3' overhangs, whether single- or double-stranded, have relatively minimal effects

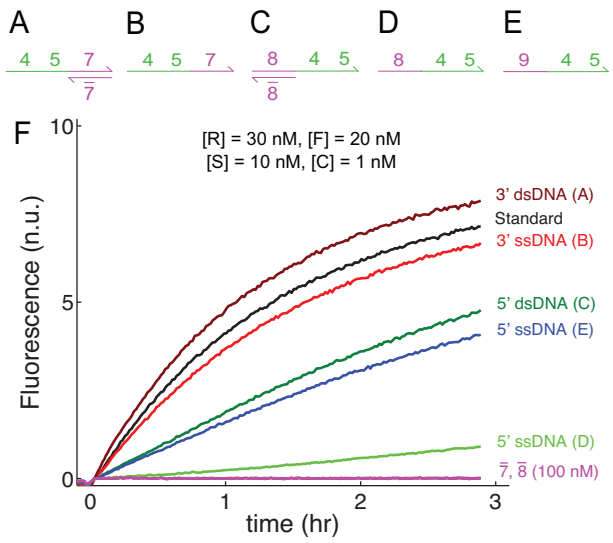


FIG. 5-4: Effects of overhangs on catalytic activity. Schematics for catalyst C with a **(A)** 3' double-stranded DNA (dsDNA) overhang, **(B)** 3' single-stranded DNA (ssDNA) overhang, **(C)** 5' dsDNA overhang, **(D)** 5' ssDNA overhang, and **(E)** 5' ssDNA overhang (mostly poly-T). **(F)** Catalytic activity of catalyst molecules with various overhangs. The magenta traces show that the $\overline{7}$ and $\overline{8}$ strands possess no catalytic activity on their own, and that the increased catalytic activity of (A) and (C) over (B) and (D) is only due to the single/double-stranded state of the overhang.

on the catalytic activity. In contrast, 5' overhangs significantly reduce catalytic activity: A double-stranded 5' overhang reduces catalytic activity by a factor of roughly 2, and a single-stranded 5' overhang reduces catalytic activity even further. The catalytic activity of the catalyst with a single-stranded 8 domain 5'-overhang is roughly $\frac{1}{8}$ that of the standard catalyst, or about a factor of 4 lower than a catalyst with a double-stranded 8 domain 5'-overhang. In contrast, the catalyst with a single-stranded 9 domain 5'-overhang exhibited catalytic activity roughly $\frac{1}{2}$ that of the standard catalyst, and comparable to the catalyst with a double-stranded 8 domain 5'-overhang.

One interpretation of these results is that the catalysts with 5'-overhangs cause a reduced rate of reaction between intermediate $I1$ and F by inhibiting the accessibility of the $\overline{3}$ domain (see Fig. 5-1B). A double-stranded 5'-overhang on C hinders F from binding due to electrostatic repulsion and steric hindrance; a single-stranded 5'-overhang on C additionally could fleetingly bind to the $\overline{3}$ domain via spurious Watson-Crick complementarities. For example, the 8 domain contains a subsequence "CCT" could fleetingly bind to the "AGG" subsequence of $\overline{3}$, despite the fact that NUPACK [28] does not predict this binding to be significant (folding at 25 °C shows $\overline{3}$ to be unbound with probability greater than 0.999

in the modified $I1$ state with the single-stranded 8 domain 5'-overhang). The extent to which spurious bindings can occur is the likely cause of the difference in catalytic activities between traces (D) and (E).

If this explanation is correct, then the effects of overhangs depend on the relative position of the overhang, with regard to the catalytic substrate S. In the current scheme, the catalyst's 5' domain acts as the recognition sequence; consequently, 5' overhangs affect the kinetics of the catalytic reaction. In a scheme where the catalyst's 5' domain acts as the toehold and the 3' domain acts as the recognition sequence, presumably 3' overhangs on the catalyst would slow the kinetics of catalysis and 5' overhangs would have minimal effect on kinetics.

Although the kinetic slowdown due to 5'-overhangs is unfortunate, it is encouraging that the catalytic reaction qualitatively works similarly even when the catalyst possesses 5'- and 3'-overhangs. This implies that such systems can be used in conjunction with biological nucleic acid inputs.

5'/3' Orientation

The dominant catalytic pathway presented in Fig. 5-1B was designed to function based on the principles of toehold exchange, which have been characterized in ref. [2]. Thus, it is expected that the same mechanism would function in a design in which all 5'/3' orientations were inverted (Fig. 5-5A). A catalytic system with an inverted 5'/3' orientation may be desirable to improve kinetics in cases where the catalyst possesses a 5' overhang but not a 3' overhang, and in certain cases to prevent spurious interactions with other strands existing in solution. Experimentally, the catalytic system with inverted 5'/3' orientation functions qualitatively and quantitatively similar to that of the original reaction (Fig. 5-5B). In these experiments, all DNA strands used were HPLC purified. See also the section titled "Effects of impurity and 5'/3' orientation on maximum turnover."

Catalyst Specificity

Regardless of whether reactions take place in a biological cell or in a test tube, in all but the simplest reaction networks there will exist a large number of molecules that may interfere with programmed reactions via spurious binding interactions. One hallmark of a

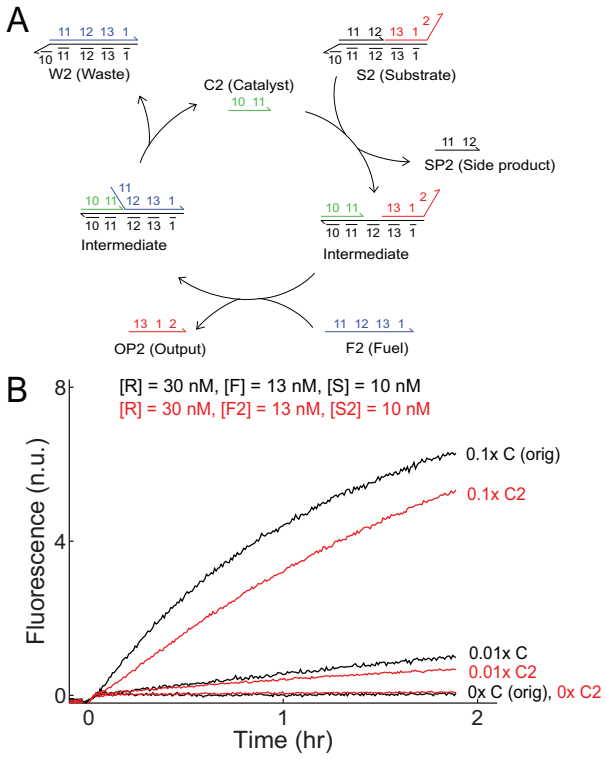


FIG. 5-5: 5'/3' inverted catalyst and substrate. **(A)** Schematic and **(B)** results. This system uses the same reporter complex as the original catalytic reaction (shown in Fig. 5-1C). Substrate $S2$ does not react significantly with reporter complex R. Note that the single-stranded 2 domain on $S2$ could hybridize to the complementary $\overline{2b}$ domain on reporter R. From there, it is possible to initiate a 4-way branch migration process that could result in the release of the fluorophore-labeled DNA. However, 4-way branch migration processes are significantly slower than 3-stranded branch migration, and the hybridization of the $2b$ domain is transient enough that this unintended pathway does not seem to be significant at our experimental conditions.

well-designed molecular amplifier is specificity: only the exactly correct input should trigger amplification. In the case of chemical amplification through DNA catalysis, specificity can be quantitatively measured as the amount of catalytic activity caused by oligonucleotides differing in sequence from the designed catalyst.

As the most stringent measure of specificity, we characterize the catalytic activities of oligonucleotides differing from the proper catalyst by only a single nucleotide. Insertions, deletions, and point "mutations" were introduced at positions 3, 7, 12, and 17 on the catalyst strand, numbered from the 5' end. Fig. 5-6ABC shows the catalytic activities of these mutant catalysts. Mutations showed strong suppression of catalytic activity, regardless of position. Insertions to the 5' of positions 3 and 7 affected the catalyst's activity less than other tested mutations; the reason for this is not yet understood.

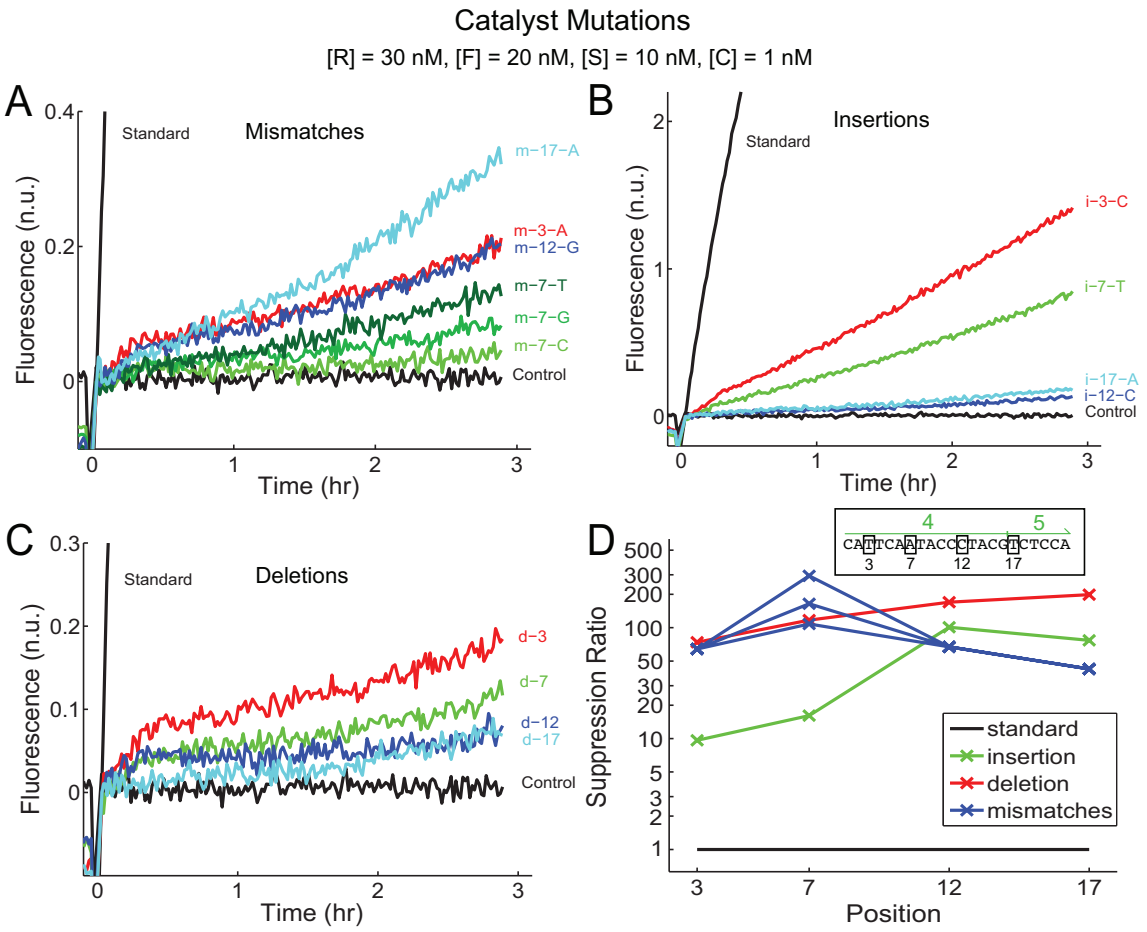


FIG. 5-6: Effects of single base catalyst C mutations. **(A)** Mismatches. The trajectory labels show the position and new identity of the mutated base. For example, "m-17-A" denotes that the 17th base of catalyst C (from the 5' end), was mutated from a Thymine to a Adenine. **(B)** Insertions. The inserted base is inserted before (5' of) the position denoted. **(C)** Deletions. **(D)** Summary of suppression by various catalyst mutations. The suppression factor is calculated as the initial slope of activity by the standard trace divided by the initial slope of activity by the mutated catalysts: $S.F. = \frac{(F_s(10\text{ min}) - F_s(0\text{ min}))/(10\text{ min})}{(F_m(3\text{ hr}) - F_m(0\text{ hr}))/(3\text{ hr})}$, where $F_m(t)$ is the fluorescence value due to a mutated catalyst at time t. **(D.inset)** The sequence of the catalyst molecule C is shown with the positions at which mutations were performed.

The suppression ratio plotted in Fig. 5-6D summarizes the effects of the tested mutations on catalytic activity. The suppression ratio is calculated by dividing the initial fluorescence slope of standard reaction (linear fit to $0 < t < 0.5$ hr data) with the fluorescence slopes of the mutant catalysts (linear fit to $0 < t < 3$ hr data). The estimation of the initial fluorescence slope of the standard catalyst differs from those of the mutant catalysts because at $t > 0.5$ hr, the standard catalytic reaction is noticeably slowed by substrate depletion and waste buildup.

All single-base mutations tested suppressed catalytic activity by at least a factor of

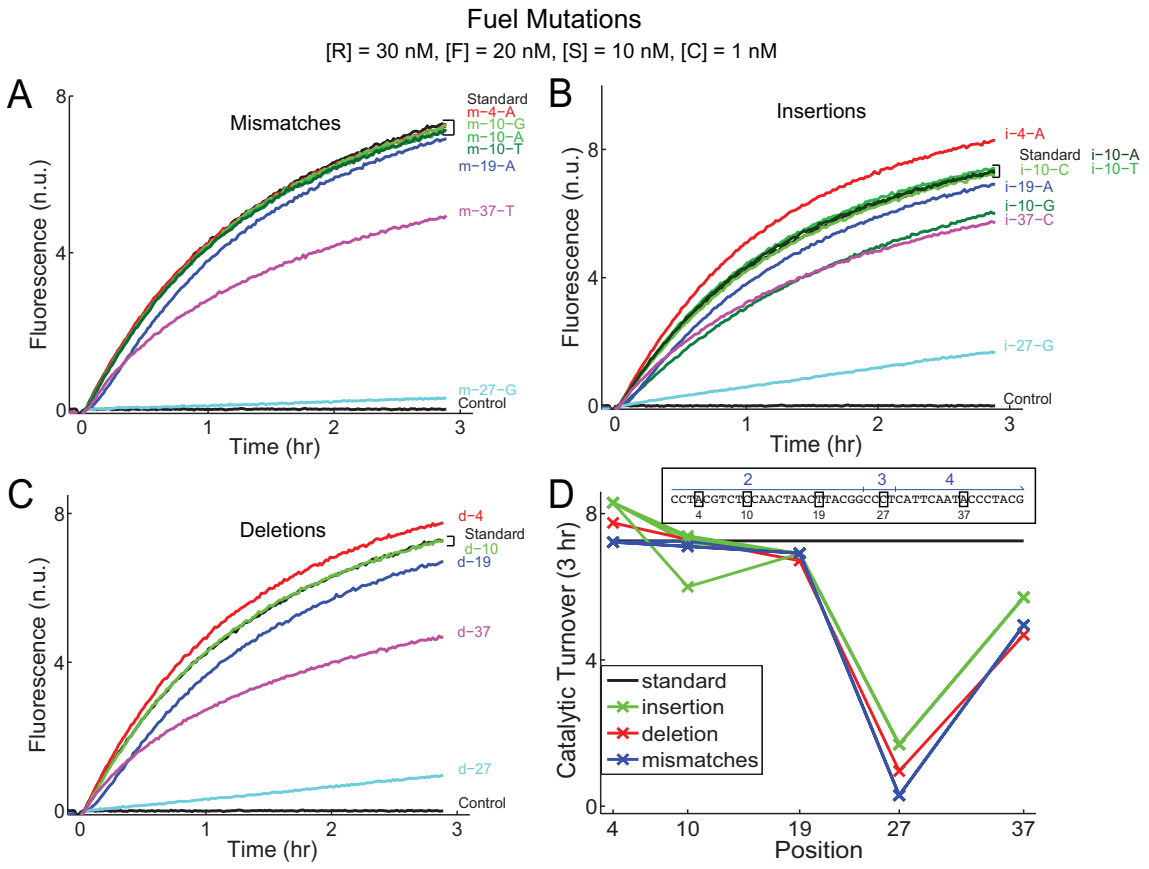


FIG. 5-7: Effects of single base fuel F mutations. **(A)** Mismatches. **(B)** Insertions. The inserted base is inserted before (5' of) the position denoted. **(C)** Deletions. **(D)** Summary of the catalytic activity using mutant fuel molecules. **(D.inset)** Sequence of the fuel molecule F and positions of mutations.

10, and the majority of them suppressed catalytic activity by 50-fold or more. This implies that in chemical reaction networks, DNA sequences can be easily designed to avoid catalytic crosstalk.

Fuel Specificity

A similar study was performed on the effects of single-base mutations on the fuel; these results are plotted in Fig. 5-7ABC. Unlike with catalysts, single-base mutations to the fuel strand yielded relatively little change in catalytic activity, except at position 27 and to a lesser extent position 37. Position 27 is in the 3 domain, which is used by the fuel F to bind intermediate $I1$; this explains why catalytic activity is sensitive to the bases in this domain. The weaker effect of fuel mutations at position 37 can likely be attributed to slower release of the catalyst from intermediate $I2$, because the branch migration process stalls at the fuel mismatch position.

The difference in specificity between the fuel and the catalyst is not accidental: the binding reaction between C and S to yield $I1$ and SP was designed to have $\Delta G^\circ \approx 0$ (Fig. 5-1B). Thus, a mutation in C destabilizes intermediate $I1$ and raises the standard free energy of the reaction as well as the standard free energies of many undocumented intermediates of the reaction. Consequently, the activation energy is raised and the kinetics are slowed. In contrast, the reaction between F and $I1$ is thermodynamically favorable (making a helix in domain 3), so a single-base mutation to F does not greatly affect the spontaneity of the $F + I1 \rightarrow I2 + OP$ or the $I2 \rightarrow W + C$ reactions.

One alternative hypothesis for explaining the difference in sensitivity between catalyst and fuel mutations is that the proposed pathway shown in Fig. 5-1B is only one of two predominant pathways–that is, $F + I1$ may yield $C + I3$, where $I3$ is a new intermediate complex composed of F, OP, and SL. Thus, a single mutation in the fuel at position 19 would bias against the shown $F + I1 \rightarrow I2 + OP$, but not affect the $F + I1 \rightarrow I3 + C$ pathway, and the second pathway would become the dominant catalytic pathway. Intermediate $I3$ would subsequently release product OP and waste W. A single mutation in the fuel at position 37 would bias against the alternative pathway, but not affect the shown $F + I1 \rightarrow I2 + OP$ pathway.

Based on our previous work on the kinetics of toehold exchange [2] we do not expect that the $F + I1 \rightarrow I3 + C$ pathway is significant: the rate constant of a toehold exchange reaction with a 4 nt invading toehold and a 6 nt incumbent toehold is at least 2 orders of magnitude slower than that with a 0 nt incumbent toehold. Thus, we expect that a catalytic reaction using a fuel with two mutations, one at position 19 and one at position 37, would have approximately the same activity as that of the fuel with a single mutation at position 37.

The simultaneous sensitivity of catalytic activity to catalyst mutations and the robustness to fuel mutations suggest the possibility of a universal fuel molecule. By designing catalysts that differ from one another by only a few bases, it could be possible to maintain reaction specificity while using the same fuel molecule to power multiple different catalytic reactions.

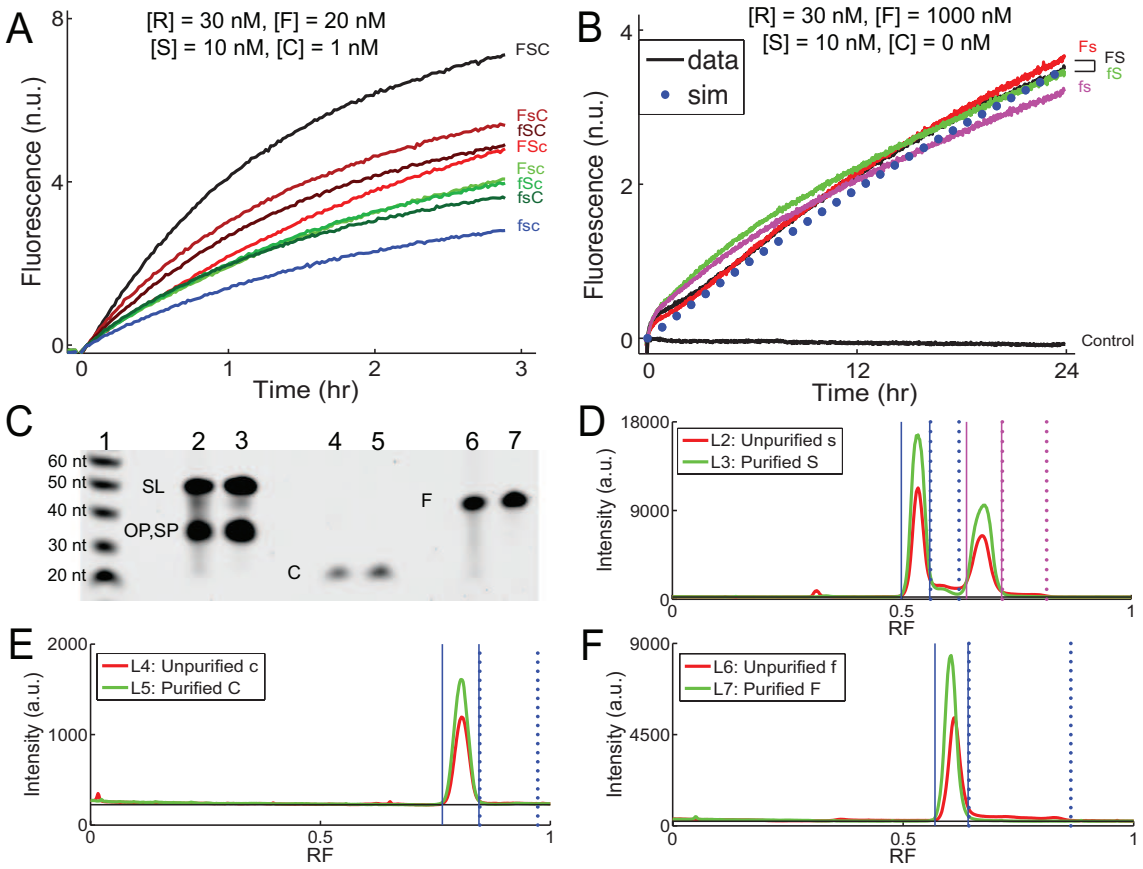


FIG. 5-8: Behavior of the catalytic reaction using fuel, substrate, and catalyst oligonucleotides with no post-synthesis strand purification. **(A)** Effects of using unpurified DNA on catalytic activity. Uppercase 'F' denotes fuel F purified commercially by HPLC, while lowercase 'f' denotes unpurified fuel. Similarly, uppercase 'S' and 'C' denote that the strands in S and the catalyst C were purified. Note that though the substrate s used unpurified DNA strands, it was still manually purified by PAGE to ensure correct stoichiometry. **(B)** Effects of unpurified DNA on the uncatalyzed reaction rate. The dotted rate shows simulation results for $k_0 = 5\ \mathrm{M}^{-1}\ \mathrm{s}^{-1}$. **(C)** Denaturing gel of purified and unpurified strands and complexes. Lanes 2 and 3 show the PAGE-purified substrate prepared from unpurified and purified OP, SP, and SL strands, respectively. This shows the degree to which truncated strands are present in PAGE-purified S and s. Lanes 4 and 5 show the unpurified and purified catalyst, respectively. Lanes 6 and 7 show the unpurified and purified fuel, respectively. **(D)** Substrate S purity. Gel band intensities are displayed in arbitrary units (a.u.). Solid blue lines denote the (arbitrarily chosen) limits of the correct-length SL strand, while dotted blue lines denote the (arbitrarily chosen) limits of the truncated SL. The black horizontal line denotes the background intensity of the gel. The integrated intensity of the correct-length SL above the background is divided by the sum of it and the integrated intensity of truncated SL to yield the correct length fraction (summarized in Table 5-4). Similarly, solid magenta lines denote the limits of correct-length OP and SP, while dotted magenta lines denote the limits of truncated OP and SP. **(E)** Catalyst C purity. **(F)** Fuel F purity.

Strand Purity Effects

In working with DNA, a large fraction of the time and cost lies not in strand synthesis, but rather in strand purification. Modern DNA strand synthesis involves a large number

of protection and deprotection reactions, which lead to inevitable errors in synthesis, in the form of various truncations, deletions, and depurinations in the synthesized oligonucleotides. While oligonucleotides with synthesis errors can be mostly removed by post-synthesis purification procedures, such as PAGE and reverse-phase HPLC, some fraction of the purified oligonucleotides still inevitably contain deletion and truncation products [26]. Furthermore, PAGE and HPLC strand purifications are expensive; the development of DNA devices robust to the oligonucleotides with synthesis errors would facilitate the widespread and practical adoption of nucleic acid-based synthetic biology.

Here, we observe the effects of using unpurified DNA oligonucleotides that contain a significant fraction of truncation products. The catalytic reaction of interest possesses two reactants (S and F), and one catalyst molecule (C). In experiments described in previous sections, all oligonucleotides involved were purchased with reverse-phase HPLC purification (by IDT). We tested the effects of replacing each of these with its unpurified counterpart, and characterized the extent to which the catalytic reaction tolerates synthesis impurities. Unpurified versions of the fuel, substrate, and catalyst are denoted by f, s, and c. Note that for s, although the three components DNA strands OP, SP, and SL are individually not purified, s is still manually purified by non-denaturing PAGE to ensure proper stoichiometry and complex formation (see Methods).

Fig. 5-8A shows the behavior of the catalytic reaction when substituting unpurified strands. The kinetics of reactions involving unpurified strands were slower than the corresponding reaction using purified strands. Furthermore, the slowdown due to using unpurified strands was cumulative–for example, using unpurified fuel and unpurified catalyst yielded slower reactions than unpurified fuel (and purified catalyst and substrate), which was in turn slower than the standard reaction with purified fuel, catalyst, and substrate. This is consistent with our results from earlier: It was shown in the previous sections that even single-base deletions can significantly impede the designed catalytic pathway. Thus, in using unpurified strands, the effective concentrations of active species are lowered.

The use of unpurified strands for fuel and substrate does not significantly change the rate constant of the uncatalyzed (leak) reaction (Fig. 5-8B). Thus, the ratio of the catalyzed

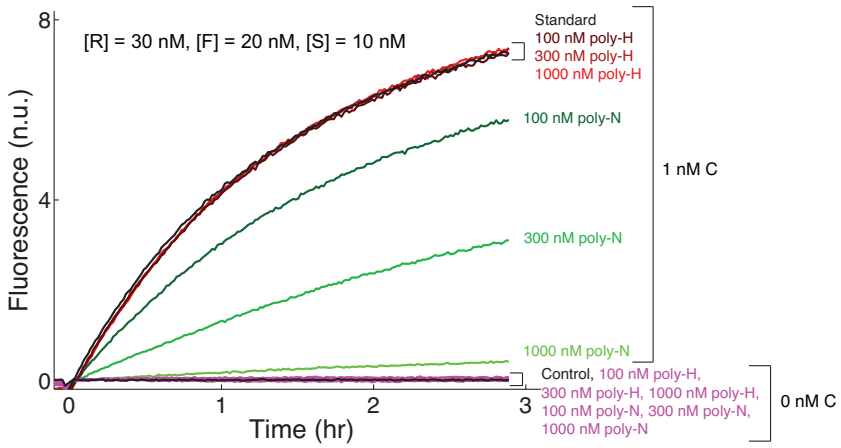


FIG. 5-9: The effects of DNA molecular noise on the kinetics of the studied catalytic reaction. Poly-N and poly-H DNA noise molecules were employed. Poly-N molecules are a mixture of different oligonucleotides, each 50 nt long, and with each base in each oligonucleotide being randomly G, C, A, or T. Poly-H molecules are similarly a mixture of 44 nt oligonucleotides, in which each base is randomly C, A, or T.

Strand	Unpurified	Purified
SL	82%	92%
OP + SP	86%	94%
C	95%	97%
F	78%	98%

Table 5-4: Summary of Correct-length Fraction

to uncatalyzed reaction rates is lower when using unpurified strands.

The degree of strand purity for both HPLC purified oligonucleotides and unpurified oligonucleotides are characterized by gel in Fig. 5-8C and quantitated in Fig. 5-8DEF. From the data in Fig. 5-8DEF, the purities are calculated and summarized in Table 5-4.

The catalytic reactions performed qualitatively similarly and quantitatively slower by no more than a factor of 2 when using unpurified DNA strands, as compared to purified strands. This robustness result implies that faulty strands can still yield robust components for DNA strand displacement-based systems.

Robustness to Background Nucleic Acid Molecules

In order for the synthetic catalysts characterized here to be easily incorporated into a complex biochemical network, it must function in an environment rife with other molecules. These "noise" molecules may interact non-specifically with the reaction substrates, catalysts, or products, changing the kinetics and thermodynamics of the intended reactions.

For example, in a cell, noise molecules include proteins, nucleic acids, and small organic molecules. In ref. [9], an autocatalytic variant of the catalytic reaction showed slower kinetics and higher leakage in the presence of cell lysate and total RNA.

In this section, we characterize the robustness properties of the catalytic reaction to DNA noise molecules. First, we tested the effects of a random mix of poly-N 50-mers (with G, C, A, or T approximately equally likely at every base position). As shown in Fig. 5-9, these random oligonucleotides never induced a false positive, in terms of catalyzing the release of product in the absence of the correct catalyst C. However, increasing concentrations of poly-N noise molecules did increasingly suppress the kinetics of the catalytic behavior; at 1 μM poly-N noise (100x that of the catalytic substrate), the catalytic activity is reduced by a factor of about 20. Contrast this result to results presented in refs. [4, 9], in which the introduction of substantial quantities of total RNA yielded relatively minor effects on the kinetics of strand displacement-based logic gates and catalytic systems. The poly-N oligonucleotides likely interact with stronger affinity for the single-stranded fuel F and catalyst C strands–in this case, the intrinsic secondary structures in biological mRNA molecules likely reduce nonspecific binding, especially because total RNA is mostly ribosomal RNA with defined structures.

For applications in synthetic chemical networks, it is desirable to possess both fast kinetics and minimal nonspecific interactions. As mentioned in the Materials and Methods, one strategy used was to minimize the frequency of G's in single-stranded domains and strands. DNA strands that do not include the G nucleotide are also unlikely to bind non-specifically to the fuel and catalyst molecules. This is consistent with experimental data (Fig. 5-9): poly-H noise molecules (in which every base is randomly C, A, or T) had no significant effect on the catalytic reaction, even at micromolar concentrations.

Characterization of a Catalyst System using a Four-Letter Alphabet

Previous experiments on the catalytic DNA reaction shown in Fig. 5-1 used domain sequences that were designed to intentionally minimize the frequency of G's in single-stranded domains. This strategy was meant to minimize spurious binding interactions between single-stranded domains, and is effective in the design of synthetic reaction networks with minimal

crosstalk interference. However, this strategy cannot be applied for systems interfacing with biology, because naturally occurring nucleic acids possess a more uniform mixture of all four nucleotides.

To investigate whether quantitatively similar catalytic reactions could be designed using a roughly equal distribution of all four nucleotides, we designed the catalytic system shown in Fig. 5-10A. The domain sequences were designed to possess minimal unwanted secondary structure, and are shown in Table 5-5. Mfold [23] predicts these sequences to possess similarly little secondary structure as the corresponding ones from the original system: At 25 °C and 11.5 mM Mg^{2+}, the predicted standard free energy (ΔG°) of the minimum free energy (mfe) structure of $C4$ was +0.12 kcal/mol (compared to +0.28 kcal/mol for C), of $F4$ was −1.34 kcal/mol (compared to −1.79 kcal/mol for F), and of $OP4$ was −0.59 kcal/mol (compared to −1.79 kcal/mol for OP) [31].

Experimentally, however, both the reporter and the catalytic reactions were significantly slower than their mostly 3-letter counterparts (Fig. 5-10BC). Fluorescence characterization of the reporter complex kinetics showed a best-fit rate constant of reaction between $OP4$ and $OR4$ of $k_{R4} = 8.2 \cdot 10^4$ M^{-1} s^{-1}. This is about a factor of 5 slower than the corresponding reporter complex designed using a 3-letter alphabet (shown in Fig. 5-1C). Similarly, the full 4-letter system also exhibited significantly slower catalytic kinetics than the original catalytic system (shown in Fig. 5-1B): 0.1x catalyst $C4$ took 10 hours to achieve 60% reaction completion with $[S4] = 30$ nM, compared to 0.1x catalyst C taking 2 hours to achieve 60% completion with one-third the substrate concentration ($[S] = 10$ nM). This implies the kinetics of the four-letter catalyst system to be roughly 15 times slower than the original catalyst system.

It was unclear why this 4-letter system performed so much slower than the original system. To continue investigating the effects of the 4-letter sequence designs, we next designed the "hybrid" catalytic system shown in Fig. 5-10D, in which the catalytic domains 63, 64, and 65 were composed of only C, A, and T, while the the output domains 51 and 52 contained all four nucleotides. At 25 °C and 11.5 mM Mg^{2+}, the predicted standard free energy of the minimum free energy structure of $F43$, $OP43$, and $C43$ were +0.02 kcal/mol,

−1.06 kcal/mol, and +0.81 kcal/mol, respectively–this is similar to the values for the other systems. The characterization of the reporter complex for this system is shown in Fig. 5-10E. The best fit rate constant for k_{R43} was $1.4 \cdot 10^5$ M^{-1} s^{-1}, which is slightly faster than that of the reporter shown in Fig. 5-10, but significantly slower than the reporter on the original 3-letter alphabet system.

The kinetics of this catalytic system (Fig. 5-10F), however, are comparable to that of the original catalyst system (and thus 15 times faster than the 4-letter system): At 10 nM substrate $S43$ concentration, 0.1x catalyst $C43$ (1 nM) achieved 60% reaction completion in $t \approx 2$ hr.

The combination of the results shown in Fig. 5-10 implies that usage of a four-letter alphabet causes a slowdown in the kinetics of toehold-mediated strand displacement and toehold exchange: Both output products $OP4$ and $OP43$ are composed of all four nucleotides and the reaction between them and their respective reporter complexes were slow.

Finally, note that the uncatalyzed reaction was significantly faster for the 4-letter system (Fig. 5-10ABC) than the hybrid system (Fig. 5-10DEF). We hypothesize that this is due to non-specific binding between the fuel $F4$ and single-stranded domain 46 on the substrate $S4$. This transient interaction results in an increased local concentration of $F4$ near $S4$, which in turn leads to a higher uncatalyzed reaction rate. In the hybrid system, the 66 domain is poly-T and unlikely to interact with $F43$.

Effects of Impurity and 5'/3' Orientation on Maximum Turnover

Modern DNA synthesis proceeds from the 3' end to the 5' end. One effect of this asymmetry is that truncations and deletions near the 5' end of DNA oligonucleotides are relatively more common than near the 3' end. This, in turn, may cause catalytic systems with different 5'/3' orientations to possess different properties. These effects will be particularly prominent when using strands not purified by HPLC or PAGE, due to the relatively higher fraction of truncated oligonucleotides.

One particular concern regarding the use of unpurified oligonucleotides is the effect of poorly synthesized fuel, even if it only exists in small quantities. For example, in the earlier section on modeling the turnover, it was hypothesized that limited turnover results from

Dom.	Sequence	Length (nt)
41	5'- TGTTACTGGCTCTGAT -3'	16
42a	5'- GACC -3'	4
42b	5'- AATGAAT -3'	7
42c	5'- ACCCGTTAC -3'	9
42	5'- GACCAATGAATACCCGTTAC -3'	20
43	5'- GAAAG -3'	5
44	5'- GCACTAAAAGTCTAC -3'	15
45	5'- CAATGTTC -3'	8
46	5'- ATAGAACATGTAGGT -3'	15
51	5'- ATAGATCCTGATAGC -3'	15
52a	5'- GAGAC -3'	5
52b	5'- CTAGCAA -3'	7
52c	5'- CCTGAAACCA -3'	10
52	5'- GAGACCTAGCAACCTGAAACCA -3'	22
63	5'- CCCTC -3'	5
64	5'- ATACAATACCC -3'	11
65	5'- TCACCATG -3'	8
66	5'- TTTTTTTTTT -3'	10

Table 5-5: Domain sequences for four-letter designs

a small fraction of F possessing defects near the 3' end. Small truncations near the 5' end of the fuel F are unlikely to have a significant effect on the catalysis of the original system [9], as represented in Fig. 5-11A. In contrast, in a catalytic system with inverted 5'/3' orientation, 5' truncations of the fuel could drastically reduce turnover (see Fig. 5-11B). For example, a fuel with a 4 nt truncation at the 5' end (Fmt in the figure) requires that the catalyst (Cm) must spontaneously dissociate 10 base pairs, rather than 6, taking hours instead of seconds.

To experimentally investigate the effects of 5'/3' orientation and purity on turnover with minimal confounding influence from sequence variation, we constructed a catalytic system using sequences that exactly mirrored the original catalytic system, with inverted 5'/3' orientation. Unlike the 5'/3' inverted system explored in Fig. 5-5, this mirrored system will possess the exact same base distribution as well as the same nearest-neighbor distribution to the original system, and is assumed to have similar thermodynamic properties as the original system, when all oligonucleotides involved are perfectly synthesized. These mirrored domains are labeled with an "m" following the original domain/strand label (e.g. Sm and $1m$, see Fig. 5-11B and Table 5-6).

Because of the relatively higher frequency of truncations and deletions near the 5' end

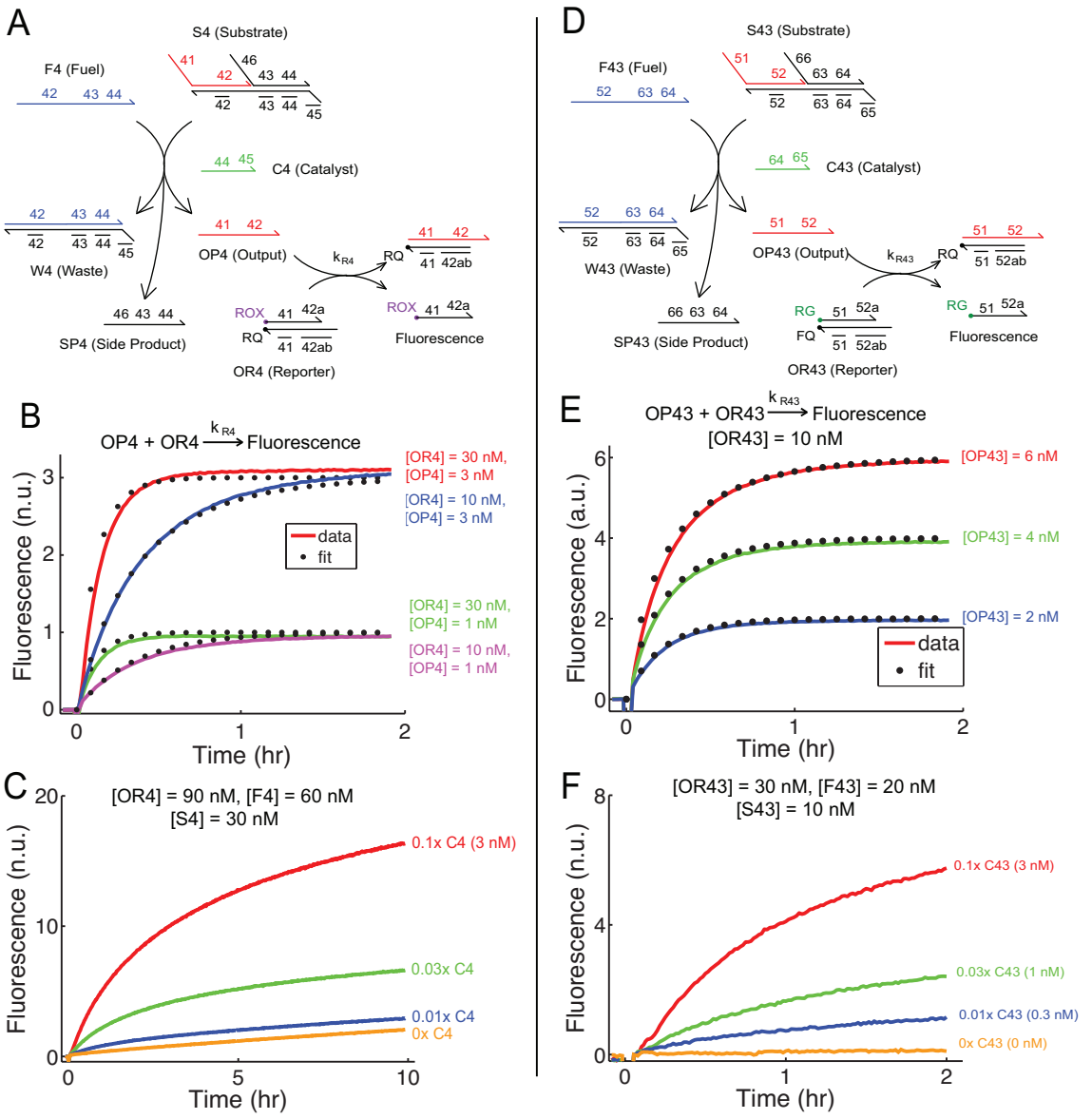


FIG. 5-10: Catalytic design using 4-letter alphabets. **(A)** Schematic of catalytic design using 4-letter base sequences. Domain sequences are given in Table 5-5. **(B)** Fluorescence characterization of the kinetics of the reporter complex. The best-fit value of k_{R4} (that produced the simulation traces shown as dotted lines) was $8.2 \cdot 10^4$ M^{-1} s^{-1}. **(C)** Fluorescence characterization of the kinetics of the 4-letter catalyst system. **(D)** Schematic of hybrid catalytic design using 4-letter base sequences in the output domains 51 and 52, and a 3-letter base sequences in the catalytic domains 63, 64, 65, and 66. Domain sequences are given in Table 5-5. "RG" on the reporter complex denotes Rhodamine Green. "FQ" on the reporter complex denotes the proprietary Iowa Black Fluorescence Quencher. **(E)** Fluorescence characterization of the kinetics of the reporter complex. The best-fit value of k_{R43} (that produced the simulation traces shown as dotted lines) was $1.4 \cdot 10^5$ M^{-1} s^{-1}. **(F)** Fluorescence characterization of the kinetics of the hybrid sequence catalyst system.

of any synthesized oligonucleotide, it is expected that the mirrored system would thus yield lower maximum turnover than the original system when unpurified strands are used.

Because oligonucleotides purified via HPLC or dual HPLC/PAGE likely still contain fuel strands with truncations and deletions, it is possible that the mirrored system would exhibit lower turnover even with these (imperfectly) purified fuel strands. The difference should be particularly striking when comparing the results from using unpurified fuel strands, in which 5' truncations are relatively common.

We tested our hypothesis by performing experiments to measure the maximum turnover of the two systems, using different fuels. These experiments are similar to those shown in Fig. 5-2, in that $[S] >> [C]$ and $[Sm] >> [Cm]$. Unlike the previous experiments on maximum turnover, however, we used unpurified strands for all but the fuel strand, in order to isolate the effects of fuel purity on the maximum turnover. (The substrates S and Sm were PAGE purified from unpurified strands.)

The first surprise was that the maximum turnover of the original catalyst system was quite low when using unpurified fuel strands (less than 10, see Fig. 5-11C). This was unexpected because it was assumed that the difference between HPLC purified strands and unpurified strands lay mostly in the absence of 5' truncated strands in the former. Previous experimentation [2] demonstrated a simplified catalytic system in which OP was removed, along with the 2 domain of F and the $\bar{2}$ domain of SL. This suggests that catalysis can occur even with very large 5' truncations in F, in which the entire 2 domain is missing (in this case, SP will be catalytically released, but not OP). One possible interpretation of the low turnover from using unpurified fuel is that the unpurified fuel strand F also possesses a significant fraction of strands with deletions near the 3' end, which the HPLC purification was successfully able to remove.

In contrast, the exact same system using HPLC purified and dual HPLC/PAGE purified fuel strand exhibited maximum turnover of over 50 over 24 hours (exact maximum turnover unknown). The results using the HPLC purified fuel strand (green trace) quantitatively agree with the results in Fig. 5-2B, implying that only high fuel strand purity is necessary for high turnover. It is unclear why the dual HPLC/PAGE purified fuel strands yielded slower catalysis kinetics than HPLC purified fuel strands.

As hypothesized, the maximum turnover of the mirrored system was significantly lower

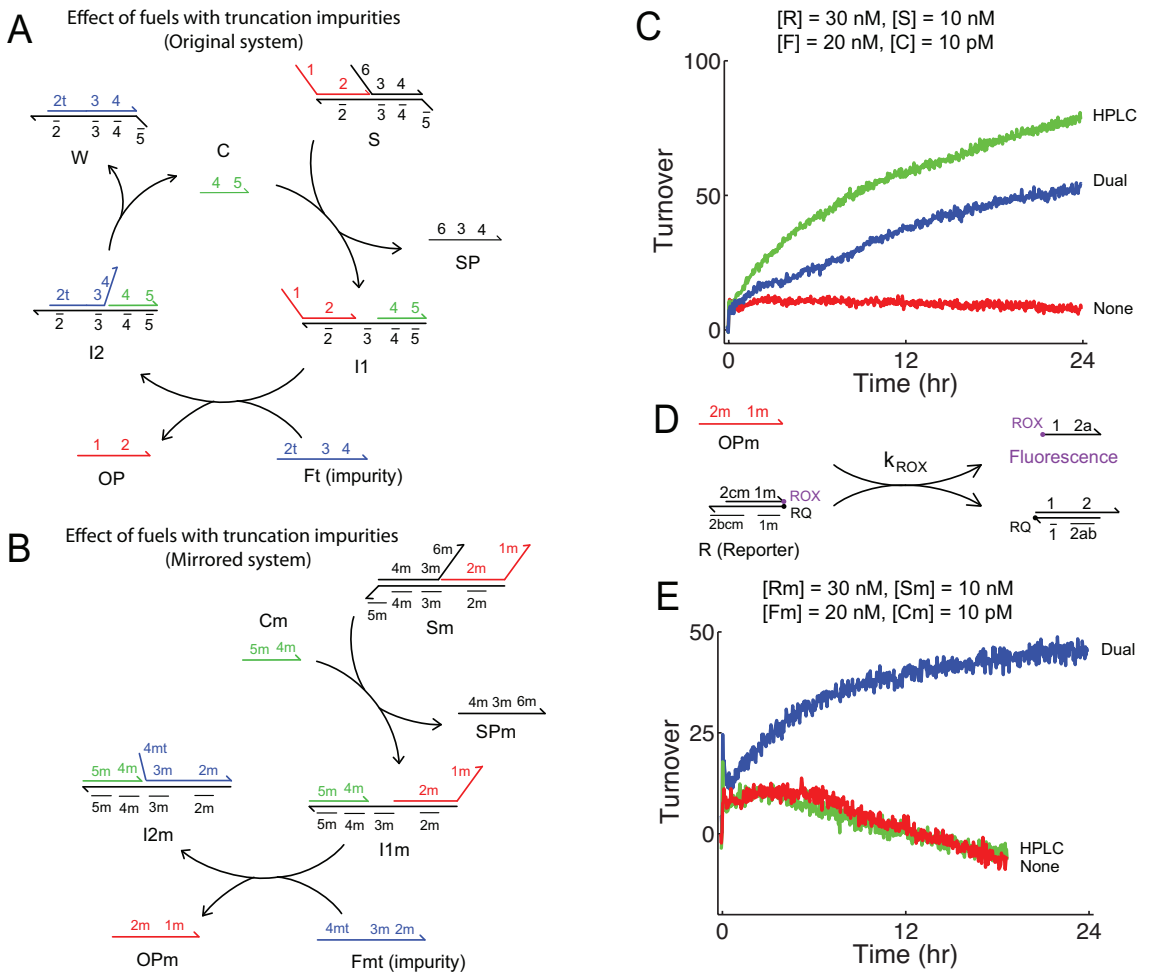


FIG. 5-11: Effects of impurity and 5'/3' orientation on maximum turnover. **(A)** Schematic of the original catalytic pathway when the fuel strand possesses an unintended 5' truncation. 5' truncated fuel strands Ft (due to synthesis errors and capping) do not hinder the designed catalytic pathway; this was shown in ref. [9]. **(B)** Schematic of the mirrored catalytic pathway when the fuel strand possesses an unintended 5' truncation. 5' truncated fuel strands Fmt are unable to quickly displace catalyst Cm from intermediate $I2m$. Catalyst cannot be regenerated to catalyze other reactions, so turnover is severely limited. **(C)** Maximum turnover of the original catalyst design. Catalyst C is unpurified, and substrate complex S is PAGE purified from unpurified strands. Fuel molecules F are unpurified ("None" trace), HPLC purified ("HPLC"), or dual HPLC/PAGE purified ("Dual"). HPLC purified and dual HPLC/PAGE purified fuels allow a maximum turnover of over 50, while unpurified fuel allows a maximum turnover of less than 10. Maximum turnover is calculated as in Fig. 5-2B: the plotted turnover is the excess fluorescence signal of an experiment with stated concentration of C over that of an experiment lacking C, divided by $[C]$. **(D)** The reporter complex for the mirror catalyst system; the strands for this reporter complex likewise mirror the sequence of the strands for the original catalyst system. **(E)** Maximum turnover measurement of the mirrored catalyst design. Catalyst Cm is unpurified, and substrate complex Sm is PAGE purified from unpurified strands. Fuel molecules Fm are unpurified ("None" trace), HPLC purified ("HPLC"), or dual HPLC-PAGE purified ("Dual"). The maximum turnover of the mirrored catalyst system using dual HPLC/PAGE purified fuel is seen to be roughly 45, while the HPLC purified and unpurified fuels allowed a maximum turnover of no more than 10. The spike near $t = 0$ is due to the subtractive nature of the method for calculating turnover, and is likely an artifact (this also exists in Fig. 5-2B). Similarly, the decline in turnover at $t > 4$ hr is likely also an artifact, due to the oligonucleotides in the uncatalyzed sample being at a slightly higher concentration (perhaps due to decreased adsorption of DNA to pipette tips) than that of the samples with catalyst.

Dom.	Sequence	Length (nt)
1m	5'- ACATCCTTTC -3'	10
2am	5'- GGCATTCAATCA -3'	12
2bm	5'- ACCTCT -3'	6
2cm	5'- GCATCC -3'	6
2m	5'- GGCATTCAATCAACCTCTGCATCC -3'	24
3m	5'- TCCC -3'	4
4m	5'- GCATCCCATAACTTAC -3'	16
5m	5'- ACCTCT -3'	6
6m	5'- TTATACTACATACACC -3'	16

Table 5-6: Domain sequences of mirrored sequences (Fig. 5-11B).

than that of the original system (Fig. 5-11EF). Using unpurified and HPLC purified fuel strands yielded maximum turnover of only 5, while dual HPLC/PAGE purified fuel strands allowed a maximum turnover of 45. The fact that the kinetics of the system using HPLC purified Fm were nearly indistinguishable from that using unpurified Fm was surprising: It was expected that HPLC purified Fm would contain only very minimal amounts of 5' truncated Fm, and thus would exhibit maximum turnover closer to that of using the dual HPLC/PAGE purified Fm.

Two general conclusions can be drawn from the results presented in this section: First, the purity of the fuel strand has a strong impact on the maximum turnover of the catalytic system, regardless of 5'/3' orientation. Given that similar maximum turnover was achieved using purified and unpurified strands for the catalyst and substrate complex, this implies that high purity of these other strands (C, SP, OP, and SL) are not as crucial for catalytic function and high turnover. Second, the 5'/3' orientation of the original catalyst system allows higher maximum turnover, because of the asymmetric nature of oligonucleotide synthesis. However, it does appear that purification of strands and complexes may be able to alleviate this intrinsic disparity between orientations.

Discussion

From the various experiments performed in this paper, several observations made of the non-covalent DNA catalytic reaction are likely to be generalizable to all strand displacement-based DNA components:

First, whenever two single-stranded domains are located in close proximity, they are likely to interfere with each other's binding to their respective complements (Fig. 5-4): Cat-

alytic activity was reduced when the catalyst possessed a 5' overhang, presumably because of its interference with the $\bar{3}$ domain. Domain interference could be due to a combination of electrostatic repulsion, steric hindrance, and non-specific binding; thus, interference is expected to be exacerbated when the lengths of the single-stranded domains increase. This observation is consistent with the slower kinetics of the previously reported allosteric catalyst [12] and hairpin-based catalytic structures [11].

Second, strand specificity can be controlled at the design level (Fig. 5-6 and 5-7): The catalytic reaction is sensitive to single-base mutations in the catalyst, but robust to mutations in the fuel (except at the toehold 3 domain). Sensitivity to differences in strand sequences is important for modularity of aqueous chemical systems, because of the large number of possible interactions that could occur in a well-mixed solution. Thus, DNA components displaying strong specificity will allow for larger and more complex synthetic chemical networks to be built. Previously, a DNA circuit with 11 components and 6 layers deep was experimentally demonstrated [4]; in theory, DNA circuits can be scaled up further and integrate more different components [27].

Third, strands do not need to be high purity in order to function (Fig. 5-8): The catalytic reaction proceeded qualitatively and quantitatively similarly with strands of less than 80% purity as with strands of over 90% purity. Impurity tolerance greatly reduces the time and cost of DNA component preparation and synthesis.

Fourth, strand displacement-based DNA components are robust to certain types of molecular noise: The DNA catalytic reaction characterized here, though sensitive to excess poly-N ssDNA, was robust to large amounts of poly-H ssDNA in solution. Just as with certain electronic modules that are robust to pink or gray but not white noise, this feature of strand displacement-based DNA systems should be considered in the design process to mitigate the effects of noise. For purely synthetic networks in which all nucleic acid molecules are designed, sequences can be chosen to take advantage of the selective noise robustness of strand displacement-based systems.

Fifth, the use of all four nucleotides in the sequence design of domains is seen to significantly slow down the kinetics of DNA toehold-mediated strand displacement, despite the

involved strands not possessing any significant predicted secondary structure. This could possibly be due to fleeting intra- and inter-molecular interactions not captured by thermodynamics, but may be revealed through simulation of DNA strand displacement kinetics at elementary step resolution [29].

Finally, the fact that DNA oligonucleotide synthesis proceeds from the 3' to the 5' end leads to an intrinsic asymmetry in the fidelity of the bases for synthetic DNA, with the 5' bases being more likely to be truncated or deleted. This, in turn, leads to a "preferred" 5'/3' orientation when designing catalytic reactions, in which lower purity strands can be used to achieved high catalytic turnover.

With the results presented herein, we remain cautiously optimistic about the future of strand displacement-based DNA circuits and networks. While integration of these constructions *in vivo* will likely require further technological breakthroughs, strand displacement-based DNA components may immediately be used to facilitate control of synthetic chemistry.

Acknowledgements. We thank Georg Seelig and David Soloveichik for insightful discussions. DYZ and EW were supported by NSF grants 0506468, 0622254, 0533064, and 0728703. DYZ is supported by the Fannie and John Hertz Foundation.

[1] SantaLucia, J. (1998) A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc. Natl. Acad. Sci.,* **95**, 1460-1465.

[2] Zhang, D. Y., Winfree, E. (2009) Control of DNA strand displacement kinetics using toehold exchange. *J. Am. Chem. Soc.,* **131**, 17303-17314.

[3] Yurke, B., Mills, A. P. (2003) Using DNA to power nanostructures. *Genet. Prog. Evol. Mach.,* **4**, 111-122.

[4] Seelig, G., Soloveichik, D., Zhang, D. Y., Winfree, E. (2006) Enzyme-free nucleic acid logic circuits. *Science,* **314**, 1585-1588.

[5] Hagiya, M., Yaegashi, S., Takahashi, K. (2006) Computing with hairpins and secondary structures of DNA. In Rozenberg, G., Back Th., Eiben, A. E., Kok, J. N., Spaink, H. P. (ed.), *Nanotechnology: Science and Computation.* Springer Berlin Heidelberg, 293-308.

[6] Frezza, B. M., Cockroft, S. L., Ghadiri, M. R. (2007) Modular multi-level circuits from immobilized

DNA-based logic gates. *J. Am. Chem. Soc*, **129**, 14875-9.

[7] Yurke, B., Turberfield, A. J., Mills, A. P., Simmel, F. C., Neumann, J. L. (2000) A DNA-fuelled molecular machine made of DNA. *Nature* **406**, 605-8.

[8] Lubrich, D., Lin, J., Yan, J. (2008) A contractile DNA machine. *Angew. Chem. Int. Ed.* **47**, 7026-7028.

[9] Zhang, D. Y., Turberfield, A. J., Yurke, B., Winfree, E. (2007) Engineering entropy-driven reactions and networks catalyzed by DNA. *Science* **318**, 1121-5.

[10] Dirks, R. M., Pierce, N. A. (2004) Triggered amplification by hybridization chain reaction. *Proc. Nat. Acad. Sci.* **101**, 15275-15278.

[11] Yin, P., Choi, H. M. T., Calvert, C. R., Pierce, N. A. (2008) Programming biomolecular self-assembly pathways. *Nature* **451**, 318-322.

[12] Zhang, D. Y., Winfree, E. (2008) Dynamic allosteric control of noncovalent DNA catalysis reactions. *J. Am. Chem. Soc.* **130**, 13921-13926.

[13] Seelig, G., Yurke, B., Winfree, E. (2006) Catalyzed relaxation of a metastable DNA fuel. *J. Am. Chem. Soc.* **128**, 12211-12220.

[14] Bois, J. S., Venkataraman, S., Choi, H. M. T, Spakowitz, A. J., Wang, Z. G., Pierce, N. A. (2006) Topological constraints in nucleic acid hybridization kinetics. *Nucleic Acids Research* **33**, 4090-4095.

[15] Turberfield, A. J., Mitchell, J. C., Yurke, B., Mills, A. P., Blakey, M. I., Simmel, F. C. (2003) DNA fuel for free-running nanomachines. *Phys. Rev. Lett.* **90**, 118102.

[16] Gartner, Z. J., Tse, B. N., Grubina, R., Doyon, J. B., Snyder, T. M., Liu, D. R. (2004) DNA-Templated Organic Synthesis and Selection of a Library of Macrocycles. *Science* **305**, 1601-1605.

[17] Aldaye, F. A., Palmer, A. L., Sleiman, H. F. (2008) Assembling materials with DNA as the guide. *Science* **321**, 1795-1799.

[18] Clo, E., Snyder, J. W., Voigt, N. V., Ogilby, P. R., Gothelf, K.V. (2006) DNA-programmed control of photosensitized singlet oxygen production. **J. Am. Chem. Soc.**, **128**, 4200-4201.

[19] Feldkamp, U., Sacca, B., Niemeyer, C. M. (2009) Dendritic DNA Building Blocks for Amplified Detection Assays and Biomaterials. *Angew. Chem. Int. Ed.* **48**, 5996-6000.

[20] Marras, S. A. E., Kramer, F. R., Tyagi, S. (2002) Efficiencies of fluorescence resonance energy transfer and contact-mediated quenching in oligonucleotide probes. *Nucleic Acids Research* **30**, e122.

[21] Gao, Y., Wolf, L. K., Georgiadis, R. M. (2006) Secondary structure effects on DNA hybridization kinetics: a solution versus surface comparison. *Nucleic Acids Research* **34**, 3370-3377.

[22] Puglisi, J. D., Tinoco, I. (1989) Absorbance melting curves of RNA. *Methods in Enzymology* **180**, 304-325.

[23] Zuker, M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nuc. Acids Res.* **31**, 3406-15.

[24] Personal communication from Bernard Yurke, Boise State University, September 2009.

[25] Bloomfield, V. A., Crothers, D. M., Tinoco, I. (2000) Nucleic Acids: Structures, Properties, and

Functions. University Science Books, Sausalito, CA.

[26] Temsamani, J., Kubert, M., Arawal, S. (1995) Sequence identity of the n-1 product of a synthetic oligonucleotide. *Nucleic Acids Research* **23**, 1841-1844.

[27] Qian, L., Winfree, E., (2009) A simple DNA gate motif for synthesizing large-scale circuits. In Goel A., Simmel F. C., Sosik, P. (ed.) *DNA Computing: 14th International Meeting on DNA Computing.* Springer Berlin Heidelberg, 70-89.

[28] Dirks, R. M., Bois, J. S., Schaeffer, J. M., Winfree, E., Pierce, N. A. (2007) Thermodynamic analysis of interacting nucleic acid strands. *SIAM Rev.* **49**, 65-88.

[29] Flamm, C., Fontana, W., Hofacker, I. L., Schuster, P. (2000) RNA folding at elementary step resolution. *RNA* **6**, 325-338.

[30] Catalysis, like other molecular processes, is stochastic, so we're actually calculating the average turnover over all the catalyst molecules present.

[31] Note that the free energy of the completely unstructured state of the DNA is defined to be 0. The mfe values reported here are as calculated by mFold [23], and correspond to the free energy of the most stable structured state, which for certain structures is less stable than unstructured.

Chapter 6: Allosteric Control

Nature is efficient; it tries not to destroy permanently something that it will reuse in the near future. Proteins are often activated and inactivated by a simple phosphate modification, and a whole array of proteins (known as kinases) deal exclusively in activating and inhibiting other proteins allosterically.

The work below demonstrates a similar level of control over the DNA catalysts presented in Chapters 3-4, via the coupled activities of allosteric DNA ligands that switch a modified DNA catalyst between its ON and OFF states. The initial idea for this work materialized in 2007, shortly after demonstrating the entropy-driven catalyst described in Chapter 3. Contrary to what's implied in the introduction below, I arrived at this idea more or less independently, taking no inspiration from similar works on ribozymes and riboswitches. Convergent thought evolution is often what happens when you have smart grad students too lazy to read papers regularly.

Initial results on a preliminary design were never published because the design experimentally exhibited a poor ON/OFF ratio (i.e. it was difficult to suppress the catalytic activity of the OFF state). The revised design, presented below, optimized the lengths of the toeholds to achieve a high ON/OFF ratio, but itself is imperfect because the switching time from the ON state to the OFF state is long.

I came up with the idea for this project, and ran the experiments. Erik Winfree helped analyze the data and revise the manuscript. Parts of this work were presented at the

14th International DNA Computing conference in Prague, Czech Republic. This work was published in full as:

Zhang, David Yu; Winfree, Erik. "Dynamic Allosteric Control of Non-covalent DNA Catalysis Reactions." *J. Am. Chem. Soc.* 130(42): 13921-13926 (2008).

Allosteric modulation of catalysis kinetics is prevalent in proteins and has been rationally designed for ribozymes. Here, we present an allosteric DNA molecule that, in its active configuration, catalyzes a non-covalent DNA reaction. The catalytic activity is designed to be modulated by the relative concentrations of two DNA regulator molecules, one an inhibitor and the other an activator. Dynamic control of the catalysis rate is experimentally demonstrated via three cycles of up- and down-regulation by a factor of over 10. Unlike previous works, both the allosteric receptor and catalytic core are designed, rather than evolved. This allows flexibility in the sequence design and modularity in synthetic network construction.

Molecular engineering using only nucleic acids has been demonstrated as a method for constructing aqueous phase logical, structural, mechanical, and catalytic elements [1–11]. Integration of these and other elements will allow the development of complex artificial biomolecular networks that may enable embedded dynamic control of cells, organisms, and synthetic nanomachines [18]. The fine modulation of kinetics is essential for this goal.

In biology, the activities of protein enzymes are finely controlled by the programmed interaction of allosteric ligands that selectively enhance or suppress the protein's catalytic behavior by inducing conformational changes. Recently, unmodified mRNAs have been observed to self-regulate via conformation change upon binding to their metabolite end-products [12–14]. In parallel, allosteric ribozymes have been constructed by joining ribozyme sequences with aptamer sequences *in vitro* [15–20]. The biological relevance of these studies was established with the discovery of similar metabolite-binding RNAs naturally occurring in bacteria [21]. Allosteric ribozymes have also been rationally designed

that catalyze other organic reactions [22].

In the works described above, the allosteric ligands that regulate the catalytic activities are small molecules such as thymine tri-phosphate (TTP). For the purposes of constructing complex synthetic biochemical networks, many different signal molecules are required for information transmission specificity among simultaneously operating network components. In order for large-scale network construction to be practical, a generalized method for engineering allosteric control elements must be developed, in which the biophysics of molecular interactions can be reliably predicted. Nucleic acids are promising candidates for the role of allosteric regulators because of their high diversity and informational content, as well as their well understood kinetics and thermodynamics. Accordingly, DNA and RNA have been designed to serve as specific regulators of allosteric ribozymes and deoxyribozymes, culminating in the construction of molecular logic gates [20, 23] and translators [24].

Here, we design and demonstrate an allosteric molecule of DNA that, in its active configuration, can catalyze a non-covalent DNA reaction. Unlike previous work, both the allosteric receptor and the catalytic core are rationally designed, rather than naturally or artificially evolved. The sequences of the DNA molecules involved are shown by domain in Table 6-1, domains being functional regions of DNA that act as a unit in binding (Fig. 6-1A).

One major advantage of pure DNA constructions is that the patterns of interactions are determined by domain complementarities and secondary structure, and largely independent of the sequences involved except insofar as sequence determines secondary structure [1–10]. For these hybridization-based DNA circuits and reactions, it is expected that designs will function for reasonable choices of sequences for each domain, because their function is based only the requirement of complementarity between complementary domains. In the sequence design of many simultaneously functioning constructions, it is important to design domains such that non-specific binding between non-complementary domains is minimal. This can be done with a number of existing sequence design software packages [25].

The allosteric catalyst (AC) presented is a single-stranded molecule of DNA that can natively adopt one of two hairpin structures, each with a different stem (Fig. 6-1B) [33]. The two states are named $AC\text{-}ON$ and $AC\text{-}OFF$ to reflect the state's catalytic activity in

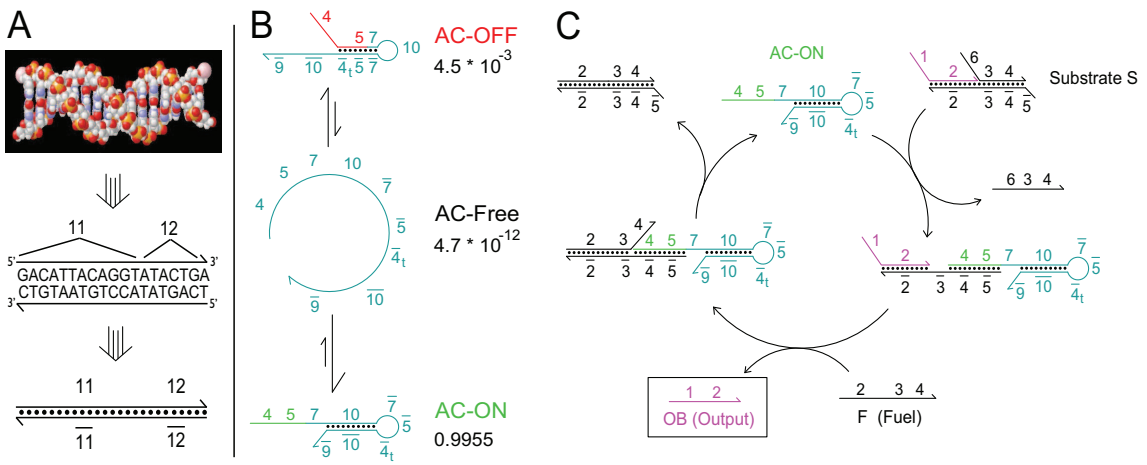


FIG. 6-1: Allosteric DNA hybridization catalyst. **(A)** DNA abstraction. The double-helix DNA molecule (top) is typically abstracted as two directional lines, one for each strand, with base identities shown (middle). Here, we abstract the DNA molecule one step further by grouping contiguous nucleotides into domains, functional regions of DNA that act as a unit in binding (bottom). Domains are labelled by numbers. Domain $\bar{x}$ is the complement of (and will hybridize to) domain x. The sequences of 11 and 12 are illustrative of the domain concept, and not in use for the allosteric catalyst design. **(B)** The allosteric catalyst (AC). There are three mechanistically important states that the AC can adopt: *AC-OFF*, *AC-free*, and *AC-ON*. Of these three, *AC-ON* is the most thermodynamically favored (due to the lengths of the hairpin stems, see Table 6-1), and *AC-free* is the least thermodynamically favored. The hairpin stem in *AC-OFF* is designed to be short enough that it can spontaneously open, causing *AC* to adopt the *AC-free* state. Then *AC* will quickly and with high probability fold into *AC-ON*. The predicted abundances of each state at equilibrium, for the sequences in Table 6-1, are shown [26]. Domain $\bar{4}t$ is a short two nucleotide domain that is complementary to the 3'-most two nucleotides of domain 4. The presence of domain $\bar{4}t$ helps ensure that the *AC-OFF* is catalytically inactive. **(C)** The catalytic cycle. The *AC* binds to the substrate *S* via domain 5 and displaces the strand 6·3·4. The newly exposed $\bar{3}$ domain allows the fuel *F* to bind, displace output *OB*, and finally displace $AC - ON$. Domains 4 and 5 on the *AC* must be single-stranded in order for the catalytic cycle to proceed.

a downstream reaction. Of the two, *AC-ON* is the thermodynamically favored state, and *AC-OFF* quickly rearranges with high probability to form *AC-ON*.

We demonstrate the *AC*'s catalytic function using the catalyst presented in [1] (Fig. 6-1C). In the catalyzed reaction, a single-stranded fuel molecule *F* reacts with a three-stranded substrate complex *S* to release a single-stranded output molecule *OB* and two waste molecules. Domains 4 and 5 on *AC* must both be single-stranded in order for the catalytic cycle to proceed [1, 3, 8]; consequently *AC-ON* is catalytically active and *AC-OFF* is not. Since *AC-ON* is the thermodynamically favored state, *AC* in isolation will be catalytically active.

The catalytic activity of the *AC* is allosterically regulated by two single-stranded species

of DNA, an Inhibitor (Inh) and an Activator (Act) (Fig. 6-2A). The catalytic activity of AC-ON is suppressed by reaction with Inh: Inh binds to AC via domain $\bar{9}$ and displaces domain 10, causing the resulting Inh:AC complex to adopt AC-OFF's hairpin. With domain 5 and part of domain 4 double-stranded, the Inh:AC complex is catalytically inactive, and thus henceforth referred to as Inh:AC-OFF. The catalytic activity is restored by reaction with Act: Act binds to Inh:AC-OFF via domain $\bar{10}$, and displaces AC in binding to domains $\bar{9}$ and $\bar{8}$, leaving AC-OFF and duplex waste product Inh:Act. AC-OFF then spontaneously rearranges to the thermodynamically favored AC-ON state. Because Inh and Act are fully complementary strands, they will preferentially bind; in equilibrium, at most one of the two is free in single-stranded form in significant concentration. Introduction of additional Inh or Act into the system (either direct intervention, or production by an upstream reaction) shifts the balance of AC-ON and Inh:AC-OFF, and can correspondingly change the rate of catalysis. We demonstrate dynamic control of production rate of OB with three cycles of up- and down-regulation.

One desirable feature of the allosteric catalyst design is that Inh and Act are both independent in sequence from the catalytic region and the downstream substrate and fuel. This allows Inh and Act to be designed as products of other reactions with minimal expected crosstalk, enabling modular kinetic control of networks.

Materials and Methods

DNA sequences and design. Strand design can be done with a number of software packages, and the proposed design should function similarly for many different sequences [25]. The sequences presented here are designed in the following manner, with no claim of being optimized for design or function efficiency. The underlying catalytic substrate and fuel sequences are the same as described in [1]; only domains 7 through 10 were designed for this work.

First, random sequences composed of only A, C, and T were generated for each unbarred domain (i.e. 2). Sequences for the complementary barred domains were constructed accordingly. Next, subsequences expected to be problematic, such as poly-G regions and

poly-A regions, were altered by hand. The remaining sequences were then concatenated as appropriate to form the DNA strands. These were folded alone and pairwise using the mFold web-server to determine possible crosstalk bindings [26]. Some bases at problematic subsequences were then changed by hand to G in the primary domains (and propagated elsewhere as appropriate) to minimize self-folding and pairwise-folding energies. Finally, the strands were checked again on mFold to ensure minimal crosstalk [26].

We chose to do sequence design "by hand" because the sequences involved were so short that this was feasible and possibly faster than specifying the parameters for DNA designs packages. For large-scale circuits, automated design would likely be necessary.

Buffer conditions. The buffer for all experiments was TE (10 mM Tris · HCl pH balanced to 8.0, 1 mM EDTA), purchased as 100x stock (Sigma-Aldrich), with 12.5 mM $MgCl_2$ added. All experiments and purifications were performed at 25 °C.

Substrate purification. DNA oligonucleotides used in this study were purchased from Integrated DNA Technologies (IDT), with HPLC purification. Where applicable, fluorophores were attached by IDT as well.

Substrate S and reporter complex OR were further purified by non-denaturing (ND) polyacrylamide gel electrophoresis (PAGE) as follows: Strands for each sample were prepared with nominally correct stoichiometry at 20 μM and annealed (only nominally correct because it is not possible to quantitate DNA strands with perfect accuracy, so a slight stoichiometric imbalance necessarily exists). A sub-stoichiometric amount (0.2x) of fuel F was added to S, allowing poorly-formed substrates to decay into products. The samples were then run on 12% ND PAGE at 180 V for 6 hours. The proper bands were cut out and eluted in 2 mL TE/Mg^{2+} buffer for 2 days. Typical yields ranged from 40% to 60%. Purified complexes were quantitated by measurement of absorbance at 260 nm using an Eppendorf Biophotometer, and concentrations were calculated using extinction coefficients for single- and double-stranded DNA predicted by nearest-neighbor models [27].

Annealing. All annealing processes were performed with an Eppendorf Mastercycler Gradient thermocycler. The samples were brought down from 95 °C to 20 °C at a constant rate over the course of 90 minutes.

Gel electrophoresis. ND PAGE was run on 12% acrylamide (19:1 acrylamide:bis), diluted from 40% acrylamide stock (Ambion). ND loading dye containing Xylene Cyanol FF (XCFF) in 50% glycerol was added to all samples, achieving final gycerol concentration of 10% by volume. Gels were run at 25 °C using a Novex chamber with external temperature bath. Gels were stained with Sybr-Gold stain (Invitrogen), and scanned with a Bio-Rad Molecular Imager. The formation gel shown in Fig. 6-2B was run at 120 V for 1 hour.

Spectrofluorimetry studies. Spectrofluorimetry studies were done using a SPEX Fluorolog-3 (Horiba) with 1.6 mL 119-004F synthetic quartz cells (Hellma). The excitation was at 588 nm, while emission was at 602 nm (optimal signal for ROX fluorophore). Slit sizes used are 2 nm for both excitation and emission monochromators for net reaction studies. Experiments shown in Fig. 6-1F and 6-2A were done with integration time of 10 seconds for every 60 second time-point. Experiment shown in Fig. 6-2C was done with integration time of 1 minute for every 5 minute time-point.

Prior to each experiment, all cuvettes were cleaned thoroughly: each cuvette was washed 15 times in distilled water, once in 70% ethanol, another 5 times in distilled water, and finally once more in 70% ethanol.

Experiments start with 1.5 mL of initial solution, and up to 51 μL of high concentration stock were added at later time points as indicated. Fluorescence values were adjusted for dilution as calculated by volume changes. The reported concentrations of species decrease 4% over the course of the experiments due to dilution. For the slit size, concentrations, and times chosen, no measurable photobleaching was observed.

Fluorescence normalization. Fluorescence (in all figures) is normalized so that 1 unit of fluorescence corresponds to 1 nM of unquenched fluorophores (Fig. 6-1E). This normalization is based on the fluorescence levels of two annealed samples: a negative control with no substrate S and a positive control with 30 nM S. In both samples, $[F] = 40$ nM and $[OR] = 60$ nM.

Carrier strands. It has been observed that DNA sticks non-specifically to pipette tips, so that serial dilutions lead to stocks more dilute than expected [1]. Unfortunately, this loss is not consistent, so we could not compensate for tip loss with additional reagent. Instead,

Dom.	Sequence	Length (nt)
1	5'- CTTTCCTACA -3'	10
2a	5'- CCTACG -3'	6
2b	5'- TCTCCA -3'	6
2c	5'- ACTAACTTACGG -3'	12
3	5'- CCCT -3'	4
4	5'- CATTCAATACCCTACG -3'	16
4t	5'- CG -3'	2
5	5'- TCTCCA -3'	6
6	5'- CCACATACATCATATT -3'	16
7	5'- TC -3'	2
8	5'- CTTGACTC -3'	8
9	5'- GTATCTAG -3'	8
10	5'- GTCTACTCCTAATG -3'	14

Table 6-1: Domain sequences

Strand	Domains	Sequence
AC	4 5 7 10 $\bar{7}$ $\bar{5}$ $\bar{4t}$ $\bar{10}$ $\bar{9}$	CATTCAATACCCTACG TCTCCA TC GTCTACTCCTAATG
		GA TGGAGA CG CATTAGGAGTAGAC CTAGATAC
Inh	8 9 10	CTTGACTC GTATCTAG GTCTACTCCTAATG
Act	$\bar{10}$ $\bar{9}$ $\bar{8}$	CATTAGGAGTAGAC CTAGATAC GAGTCAAG
F	2 3 4	CCTACGTCTCCAACTAACTTACGG CCCT
		CATTCAATACCCTACG
OB	1 2	CTTTCCTACA CCTACGTCTCCAACTAACTTACGG
SB (part of S)	6 3 4	CCACATACATCATATT CCCT CATTCAATACCCTACG
LB (part of S)	$\bar{5}$ $\bar{4}$ $\bar{3}$ $\bar{2}$	TGGAGA CGTAGGGTATTGAATG AGGG
		CCGTAAGTTAGTTGGAGACGTAGG
OF (part of OR)	1 2a	CTTTCCTACA CCTACG
OQ (part of OR)	$2\bar{b}$ $2\bar{a}$ $\bar{1}$	TGGAGA CGTAGG TGTAGGAAAG

Table 6-2: Strand sequences

we introduced into all dilute stocks (1 μM and below) a non-reactive 20 nt poly-T "carrier" strand, at a concentration of 1 μM. Because pipette tip loss is non-specific, the majority of DNA loss would be of the carrier strand, and serially diluted stocks are only slightly more dilute than expected. Poly-T strands have minimal influence the reactions of other DNA molecules [1].

Results and Discussion

State changing. The use of partially complementary nucleic acid molecules as regulators to effect state changes has been demonstrated for RNA switches [18, 19], DNA motors [4, 32], and DNA catalysts [1, 5, 7]. Their kinetics are generally rate limited by the second-order initiation reaction, which can be adjusted based on the lengths and sequences

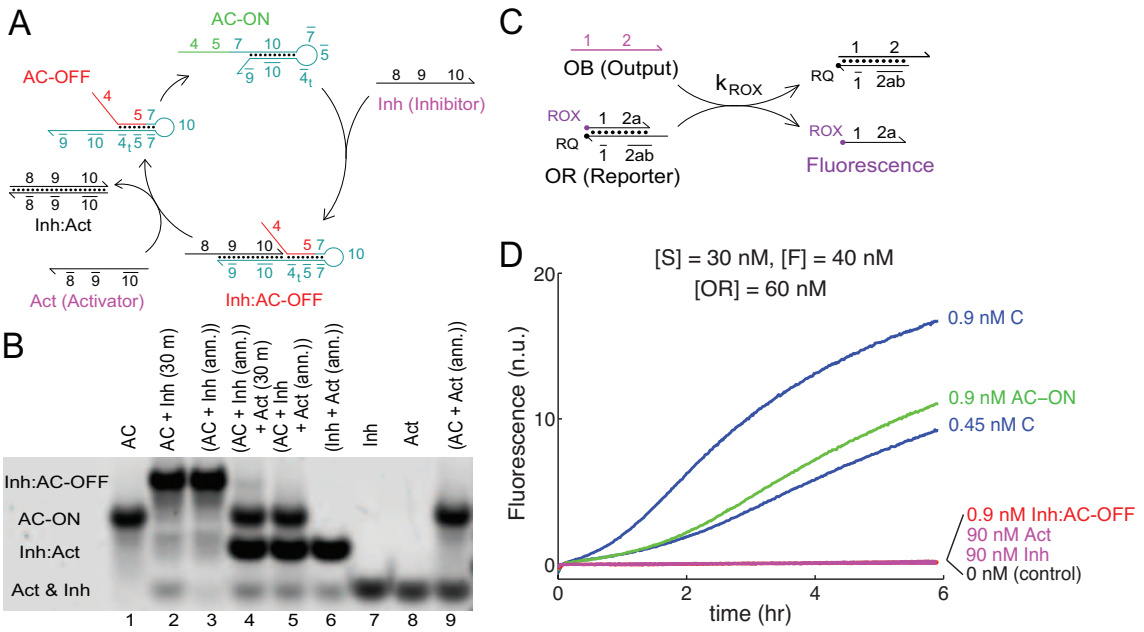


FIG. 6-2: State changing of the AC. **(A)** State-changing by addition of inhibitor (Inh) and activator (Act). Free Inh binds to AC-ON to form the Inh:AC-OFF complex. Free Act binds to Inh:AC-OFF to release AC-OFF and duplex waste product Inh:Act. AC-OFF then spontaneously converts to AC-ON. **(B)** Analysis by PAGE (12% native gel) of state changing. Here, $[AC] = 200$ nM, $[Inh] = 250$ nM, and $[Act] = 300$ nM. "(ann.)" denotes that species were annealed; "(30 m)" denotes that the reaction proceeded for 30 minutes. Lanes 1 through 5 show slight smearing because the allosteric catalyst can dynamically switch between its two states over the course of the gel running. **(C)** The production of OB is quantitated via stoichiometric reaction with reporter complex OR to yield increased fluorescence. ROX denotes the carboxy-X-rhodamine fluorophore (attached to the DNA molecule via an NHS ester), and RQ denotes the Iowa Black Red Quencher. Domain 2 is subdivided into 2a, 2b, and 2c; 2ab consists of 2a and 2b (Table 6-1). The concentration of the OR reporter complex was always in excess of S to ensure that reporting delay time was approximately consistent. **(D)** Catalytic activity of the allosteric catalyst. Fluorescence (in all figures) is normalized so that 1 normalized unit (n.u.) of fluorescence corresponds to 1 nM of unquenched fluorophore-labelled strand 1·2a. Various reagents were added at $t \approx 0$. Serving as a control, catalyst C from [1] has sequence 4·5. Inh:AC-OFF was prepared by annealing AC with a 3x excess (2.7 nM final concentration) of Inh.

of of the nucleic acids involved. In the limit of long single-stranded domains, these rates approach the hybridization rate of complementary DNA strands.

Although it is desirable for the kinetics of state switching to be as fast as possible, the rate constants need not be quite as high as those of the underlying catalytic mechanism, because high concentrations of regulators Inh and Act can be used to induce a higher rate. The efficacy of state-switching by Inh and Act is demonstrated by ND PAGE in Fig. 6-2B. In isolation, AC forms a well-defined band, indicating the predominance of AC-ON over AC-OFF. Upon addition of slight excess (1.25x) of Inh, almost all of AC migrates in the

Inh:*AC-OFF* band after 30 minutes of reaction, indicating that reactions are fast on this time scale. Similarly, addition of a slight excess (1.5x) of *Act* displaces almost all of *Inh* from *Inh*:*AC-OFF* within 30 minutes, forming the duplex band *Inh*:*Act* and regenerating the *AC-ON* band.

The time scale of the rearrangement from *AC-OFF* to *AC-ON* can be estimated from DNA thermodynamics and known rate constants: The rate at which *AC-OFF* opens to form *AC-Free* should be similar to the rate at which two complementary strands with similar binding energy dissociate. This is calculated to be $k_f e^{-\Delta G/RT}$, where k_f is the rate constant of DNA hybridization and ΔG is the energy of binding. Recent reported measurements of this value range from 10^5 /M /s to 10^6 /M /s [28]. However, our own experiments in TE / Mg^{2+} buffer with DNA molecules of similar nucleotide compositions in similar buffer conditions suggest a hybridization rate of $k_f = 3 \cdot 10^6$ /M /s (data not shown). Using the latter value and *AC-OFF*'s free energy (-12.3 kcal / mol, as predicted by mFold), the opening rate is estimated to be $3 \cdot 10^{-3}/s$, giving a timescale of about 5 minutes. Using lower values of k_f, the opening rate is accordingly decreased.

The unstructured *AC-Free* state then very quickly adopts either the *AC-ON* or the *AC-OFF* fold. Since there are more bases in the stem of *AC-ON* than in *AC-OFF*, *AC-Free* is more likely to fold into *AC-ON*. This implies that the expected time for rearrangement is no more than 10 minutes. In contrast, similar arguments predict that *AC-ON* (-15.5 kcal / mol) opens to the *AC-Free* state on a timescale of about 18 hours.

Catalytic activity of the allosteric catalyst. The catalytic activity of *AC* in its various states is inferred by production of the output molecule *OB*. This is measured using a fluorescent reporter complex *OR*: *OB* reacts stoichiometrically with *OR* to release an unquenched fluorescent strand (Fig. 6-2C). Before *OB* is released from the catalytic substrate *S*, it does not react with *OR*.

Fig. 6-2D shows the catalytic turnover of *AC-ON*, *Inh*:*AC-OFF*, and control catalyst *C*. The suppression of catalysis in *Inh*:*AC-OFF* is almost complete; the fluorescence levels of the reaction with 0.9 nM *Inh*:*AC-OFF* is indistinguishable from that of the control experiment with neither *C* nor *AC*. On the other hand, *AC-ON* retains catalytic activity:

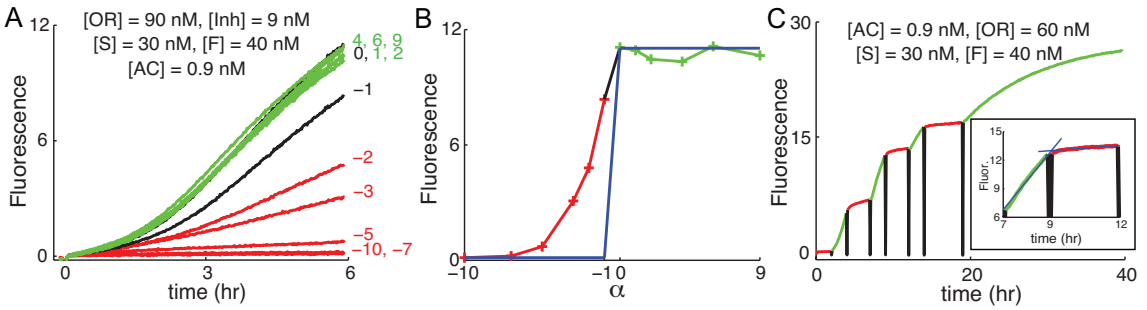


FIG. 6-3: Allosteric catalyst behavior. **(A)** Dependence of catalytic behavior on the balance of $[Act]$ and $[Inh]$. $[S]$ = 30 nM = 1x, $[F]$ = 40 nM = 1.3x, $[AC]$ = 0.9 nM, $[Inh]$ = 9 nM. Various amounts of Act were added at $t = 0$ min, with the number label shown being the value of α, the stoichiometric excess of Act. Red shows where the AC is expected to be OFF (Inh:AC-OFF); green shows where AC is expected to be ON (AC-ON). **(B)** Sigmoidal activation curve. The total catalytic activity over 6 hours (from (A)) is plotted against α. Blue trace denotes the expected behavior when AC, Inh, and Act are in equilibrium. **(C)** Dynamic switching. Initially, 0.9 nM AC is annealed with 2.7 nM Inh. Reagents were added to cause concentration changes as follows: 20 nM Act at $t = 2$ hr, 40 nM Inh at $t = 4$ hr, 60 nM Act at $t = 7$ hr, 80 nM Inh at $t = 9$ hr, 100 nM Act at $t = 12$ hr, 120 nM Inh at $t = 14$ hr, 140 nM Act at $t = 19$ hr. The fluorescence level was adjusted for dilution by multiplying by the dilution factor where appropriate. **(C.inset)** Rate fitting for observed activity between 7 and 12 hours. The production rate of OB was fit to be 3.0 nM / hr for AC-ON, and 0.18 nM / hr for Inh:AC-OFF (blue traces).

0.9 nM catalyze the release of over 10 nM of OB over 6 hours. This represents a 40% slowdown compared to control catalyst C, which is signficant but not critical. The slowdown is suspected to be due to steric hindrance and/or electrostatic repulsion: Domain 9 is close to the active catalytic domains 4 and 5, and may interfere with the binding of those domains to substrate S.

Inh and Act are both independent in sequence from the catalytic region and the downstream substrate and fuel, and thus are expected not to show any catalytic activity at all. This is indeed the case, as even 90 nM of Act or Inh alone failed to cause any observable reaction.

Activation of catalytic activity. The catalytic activity of the AC changes according to the balance of Inh and Act. Define α to be the stoichiometric excess of Act over Inh relative to AC: $\alpha = (\frac{[Act]_{total} - [Inh]_{total}}{[AC]_{total}})$. When AC and its allosteric effectors are at equilibrium, the catalytic activity can be divided into three regimes based on the value of α. For $\alpha \leq -1$, nearly every molecule of AC is expected to be bound in the Inh:AC-OFF state, and catalytic activity should be minimized. For $-1 < \alpha < 0$, some but not all molecules

of AC are bound in Inh:AC-OFF, and the catalytic activity is at an intermediate value depending on the exact value of α. For $\alpha \geq 0$, nearly every molecule of AC is free in the AC-ON state, and the catalytic activity should be maximized (Fig. 6-3B).

When the AC is used as part of a larger dynamic network, equilibrium is not maintained during function. To characterize the dynamic behavior of AC's activation, we pre-annealed the AC with a 10x excess of Inh, and then added various amounts of Act (Fig. 6-3A). The net effect over 6 hours of reaction is summarized in Fig. 6-3B.

The major difference between these experiments, and the cases when Inh is pre-annealed with Act, is that the AC shows partial catalytic activity when $\alpha < -1$. In contrast, when $\alpha > 0$, the catalytic activity is very close to that of AC-ON. The latter suggests that switching from Inh:AC-OFF to AC-ON is relatively fast, while the former suggests that the reverse process is relatively slow: When $\alpha < -1$, the amount of Act added is lower than the amount of free Inh, but nonetheless some Act will react with Inh:AC-OFF rather than free Inh. The resulting AC-OFF rearranges quickly into the active AC-ON state, remaining in solution until it reacts once again with free Inh. This last step being slow on the timescale of the experiment explains the continued catalytic activity for $\alpha < -1$.

The relatively slower kinetics of turning the AC off with Inh is suspected to be due to non-specific binding interactions between the $\bar{9}$ domain and the 4, 5, and 7 domains. The close proximity between these domains means that the local concentrations are very high, so that one- and two- base pairings, which normally do not form, may exist fleetingly. Inaccessibility of the $\bar{9}$ domain is expected to decrease the rate of state changing, because the binding of Inh to AC-ON is initiated and rate limited by the binding of Inh's 9 domain to AC-ON's $\bar{9}$ domain. In contrast, Inh:AC-OFF does not have a 3' overhang that non-specifically interacts with Inh's 8 domain, so the rate of switching from Inh:AC-OFF to AC-ON is not reduced.

Dynamic switching. The AC can be dynamically switched between the its catalytically active (AC-ON) and inactive (Inh:AC-OFF) states. Fig. 6-3C shows the production of OB, as modulated by addition of Inh or Act at various time points (shown by black spikes). There is a clear difference in the slopes between when Act was in excess (shown in

time (hr)	state	OB production rate (nM / hr)
0 - 2	OFF	0.00
2 - 4	ON	3.5
4 - 7	OFF	0.24
7 - 9	ON	3.0
9 - 12	OFF	0.18
12 - 14	ON	1.6
14 - 19	OFF	0.08

Table 6-3: Catalytic activity based on state

green) and when Inh was in excess (shown in red).

The rates of production of OB between reagent additions are fitted (Fig. 6-3C.inset), and the results are summarized in Table 6-3. The first 20 minutes of data after reagent additions were not used for these fits. The decrease in production rate of OB over time can be attributed to decreasing substrate (S) and fuel (F) concentrations, and to a lesser extent accumulated DNA loss to pipette tips (see Materials and Methods: Carrier Strands).

The kinetics of switching AC-ON to Inh:AC-OFF is seen to be significantly slower than that of the reverse (time scale of approx. 30 min, as opposed to less than 5 min), supporting the hypothesis that switching the AC off is significantly slower than switching it on. Given the excess concentrations of Inh and Act at various times, we estimate the second order reaction rate between AC-ON and Inh to be roughly $3 \cdot 10^4/M/s$, and that between Inh:AC-OFF and Act to be greater than $2 \cdot 10^5/M/s$.

Conclusions. We have demonstrated allosteric control of non-covalent DNA-based catalysis based on alternative hairpin folds. Two molecules controlled the catalytic activity, a non-competitive inhibitor, and an activator that competitively inhibits the inhibitor. The activator and inhibitor have the same form (single-stranded DNA molecules of similar size) as the released product molecule of the catalytic reaction. Thus, they can be the products of other synthetic biomolecular reactions, released dynamically through the course of operation.

The hybridization-based mechanism by which the allosteric catalyst changes physical and catalytic states is simple and well-understood [32]. This method of state control can, in principle, be extended to apply to other nucleic acid based constructions, such as catalytic networks, logic gates, structural assembly, etc.

The preferential binding of the inhibitor *Inh* to activator *Act* causes a sigmoidal non-linearity between catalytic activity and the amount of activator added, given a constant inhibitor concentration (Fig. 6-3B). Similar mechanisms have been biologically observed in kinase cascades to produce ultrasensitivity, filtering out biochemical "noise" [31]. From the network construction point of view, it has been proposed that such thresholding behavior can be used to convert between analog and digital representations of information in chemical reaction networks [29]. Recently, thresholds constructed from the competitive binding of mRNA and DNA have been used to build bistable transcriptional networks [30]. Thus, the current work may be used to engineer enzyme free reaction networks robust to noise.

In previous work on allosteric ribozymes and riboswitches, the identities of the allosteric effectors are determined by aptamer sequences involved, and cannot be easily modified. In this work, the allosteric effectors are molecules of DNA whose sequences can be easily programmed: they are independent of the catalytic core, as well as independent of the output molecule. Thus, many different allosteric effectors should be able function simultaneously without significant crosstalk.

Acknowledgements. The authors were supported by NSF grants 062254 and 0728703 to EW. DYZ is supported by the Fannie and John Hertz Foundation.

[1] D. Y. Zhang, A. J. Turberfield, B. Yurke, E. Winfree, *Science* **318**, 1121 (2007).

[2] G. Seelig, B. Yurke, E. Winfree, *J. Am. Chem. Soc.* **128**, 12211 (2006).

[3] G. Seelig, D. Soloveichik, D. Y. Zhang, E. Winfree, *Science* **314**, 1585 (2006).

[4] B. Yurke, A. J. Turberfield, A. P. Mills, F. C. Simmel, J. L. Neumann, *Nature* **406**, 605 (2000).

[5] A. J. Turberfield, J. C. Mitchell, B. Yurke, A. P. Mills, M. I. Blakey, F. C. Simmel, *Phys. Rev. Lett.* **90**, 118102 (2003).

[6] E. Winfree, F. Liu, L. A. Wenzler, N. C. Seeman, *Nature* **394**, 539 (1998).

[7] R. M. Dirks, N. A. Pierce, *Proc. Nat. Acad. Sci.* **101**, 15275 (2004).

[8] P. Yin, H. M. T. Choi, C. R. Calvert, N. A. Pierce, *Nature* **451**, 318 (2008).

[9] H. Yan, X. Zhang, Z. Shen, N. C. Seeman, *Nature* **415**, 62 (2002).

[10] B. Ding and N. C. Seeman, *Science* **314**, 138 (2006).

[11] P. Rothemund, *Nature* **440**, 297 (2006).

[12] A. Nahvi, N. Sudarsan, M. S. Ebert, X. Zou, K. L. Brown, R. R. Breaker, *Chemistry & Biology* **9**, 1043 (2002).

[13] A. S. Mironov, I. Gusarov, R. Rafikov, L. E. Lopez, K. Shatalin, R. A. Kreneva, D. A. Perumov, E. Nudler, *Cell* **111**, 747 (2002).

[14] J. E. Barrick, K. A. Corbino, W. C. Winkler, A. Nahvi, M. Mandal, J. Collins, M. Lee, A. Roth, N. Sudarsan, I. Jona, J. K. Wickiser, R. R. Breaker, *Proceedings of the National Academy of Sciences* **101** 6421 (2004).

[15] J. Tang, R. R. Breaker, *Chemistry & Biology* **4**, 453 (1997).

[16] H. Schwalbe, J. Buck, B. Furtig, J. Noeske, and J. Wohnert, *Angewandte Chemie Int. Ed.* **46**, 1212 (2007).

[17] A. D. Ellington, J. W. Szostak, *Nature* **355**, 850 (1992).

[18] F. J. Isaacs, D. J. Dwyer, C. M. Ding, D. D. Pervouchine, C. R. Cantor, J. J. Collins, *Nature Biotech.* **22**, 841 (2004).

[19] F. J. Isaacs, D. J. Dwyer, J. J. Collins, *Nature Biotechnology* **24**, 545 (2006).

[20] R. Penchovsky, R. R. Breaker, *Nature Biotech.* **23**, 1424 (2005).

[21] W. Winkler, A. Nahvi, R. R. Breaker, *Nature* **419**, 952 (2002).

[22] S. Amontov, A. Jaschke, *Nucleic Acids Research* **34**, 5032 (2006).

[23] M. N. Stojanovic, T. E. Mitchell, D. Stefanovic, *J. Am. Chem. Soc.* **126**, 3555 (2002).

[24] J. J. Tabor, M. Levy, A. D. Ellington, *Nucleic Acids Research* **34**, 2166 (2006).

[25] R. M. Dirks, M. Lin, E. Winfree, N. A. Pierce *Nucleic Acids Res.* **32**, 1392 (2004).

[26] M. Zuker, *Nucleic Acids Res.* **31**, 3406 (2003).

[27] J. D. Puglisi, I. Tinoco, *Methods in Enzymology* **180**, 304 (1989).

[28] Y. Gao, L. K. Wolf, R. M. Georgiadis, *Nucleic Acids Research* **34**, 3370 (2006).

[29] M. O. Magnasco, *Phys. Rev. Lett.* **78**, 1190 (1997).

[30] J. Kim, K. S. White, E. Winfree, *Molecular Systems Biology* **2**, 68 (2006).

[31] J. E. Ferrell, *Trends in Biochemical Sciences* **21**, 460 (1996).

[32] B. Yurke, A. P. Mills, *Genetic Programming and Evolvable Machines* **4,** 111 (2003).

[33] The design of the *AC* must consider the possibility of pseudoknot formation due to necessary sequence self-complementarity. For the presented design, the hairpin stem regions are short enough that we do not expect significant pseudoknot formation.

Chapter 7: Digital Nucleic Acid Concentration Sensors

One potential application of the entropy-driven DNA catalyst system described in Chapter 3 is DNA and RNA detection and quantitation. Fortunately or unfortunately, the state of the art for standard nucleic acid detection using quantitative PCR (qPCR) is already very good, with sensitivity in the atto- to zepto-mole range. Considering that the best detection demonstrated using the entropy-driven DNA catalyst networks was 1 femtomole (1 pM $\cdot$ 1 mL), we lag the current state of the art by 3 to 6 orders of magnitude for standard DNA and mRNA detection. However, qPCR requires both enzymatic function and thermal cycling, and is not well suited for sensing DNA concentrations in biological settings (temperatures, salinities, etc.). This means, among other things, that qPCR cannot be integrated with synthetic biochemical nucleic acid control circuits.

In order for engineered nucleic acid reaction networks to properly interface with biological nucleic acids, a reliable and robust method of sensing oligonucleotide concentrations is necessary. While the entropy-driven DNA catalyst system can amplify nucleic acid signals to allow more sensitive detection, variations in the kinetics of such designed reactions prevent them from being accurate quantitation devices. In other words, the entropy-driven DNA system is an analog amplification mechanism, and the gain of the system varies by design and cannot be precisely predicted.

In the work below, I demonstrate a circuit that digitally compares the concentration of two oligonucleotides of independent sequence. This mechanism is then modularly coupled to the entropy-driven DNA catalyst mechanism to demonstrate amplified thresholded quantitation. The detection sensitivity demonstrated is only about 1 pmol, but this can be improved through system optimization. This work, like all other works in this thesis, is based only on the biophysics of DNA hybridization and strand displacement, and thus enzyme-free and isothermal.

I came up with the initial idea for this work in 2006 shortly after co-authoring the nucleic acid logic gates paper with Georg Seelig, David Soloveichik, and Erik Winfree. Initial experimentation on this project generally did not work, and this project was shelved for about 2 years. In 2008, I re-examined this system and realized that my toeholds from my earlier design were likely too short, and their binding was insufficient to overcome the entropy cost of associating one extra molecule at the experimental concentrations. I then redesigned the system, and proceeded to do the experiments for this project through the course of 2008 and 2009.

This work was submitted for publication on April 9, 2010 as:

Zhang, David Yu. "Digital Sensors for Comparing Oligonucleotide Concentrations."

The predictability of nucleic acid interactions [1, 2] has enabled the construction of dynamic nucleic acid devices that autonomously move [3, 4], perform logical computation [5, 6], or control self assembly [7–10]. The behavior of many of these designs depend on the relative concentrations of oligonucleotides; embedded digital control circuits that regulate oligonucleotide concentrations could thus enable the programming of more complex behaviors [5, 6, 11]. As a first step, it is necessary to develop a modular nucleic acid system that accurately senses and evaluates the relative concentrations of oligonucleotides in solution. Here, a reaction network that digitally and dynamically compares the concentrations of two oligonucleotides of independent sequence is experi-

mentally demonstrated. The system is modularly cascaded with a rationally designed DNA catalysis reaction for signal amplification and improved sensitivity. This embedded concentration quantitation network represents an important milestone in the construction of synthetic biochemical circuits.

The systems presented in this paper employ a novel reaction mechanism that I call cooperative hybridization. Cooperative hybridization is a rationally designed series of nucleic acid hybridization, branch migration, and dissociation events that ensures the simultaneous and stoichiometric reaction of two oligonucleotides of independent sequence (Fig. 7-1A). Two single-stranded oligonucleotides, $T1$ and $T2$, can hybridize simultaneously to multi-stranded DNA device $D1$, but individual binding is unfavorable. When both $T1$ and $T2$ are present, the reaction follows one of two pathways. In the first pathway, $T1$ initially hybridizes via domain 1 to $D1$, and then displaces domain 2 through branch migration to form intermediate I. Next, $T2$ hybridizes via domain 4 to intermediate I, and displaces domain 3 through branch migration. Product $P1$ is released, and the stoichiometrically bound $T1$ and $T2$ are sequestered in the waste product W. In the other pathway, $T2$ reacts with $D1$ first to form intermediate J, which subsequently reacts with $T1$ to release $P1$ and W.

This mechanism is cooperative because the individual binding of $T1$ or $T2$ to $D1$ is thermodynamically unfavorable, and quickly reversible: domains 1 and 4 are short (8 nt) and the free energy gain derived from the hybridization of either domain to its complement is insufficient to overcome the configurational entropic loss of colocalizing an extra DNA strand at experimental concentrations. On the other hand, the hybridization of $T1$ or $T2$ to J or I, respectively, is thermodynamically favorable because of the release of $P1$. The net reaction of $T1$, $T2$, and $D1$ to form W and $P1$ is also thermodynamically favorable because the hybridization of both domains 1 and 4 to their respective complements forms 16 new base pairs, which is enough to overcome the net entropy loss of of the reaction at experimental concentrations. For different operating concentrations, the lengths/sequences of domains 1 and 4 may need to be altered (see Text 7-S1).

To observe the kinetics of this mechanism, reporter complex R is used (Fig. 7-1B). R

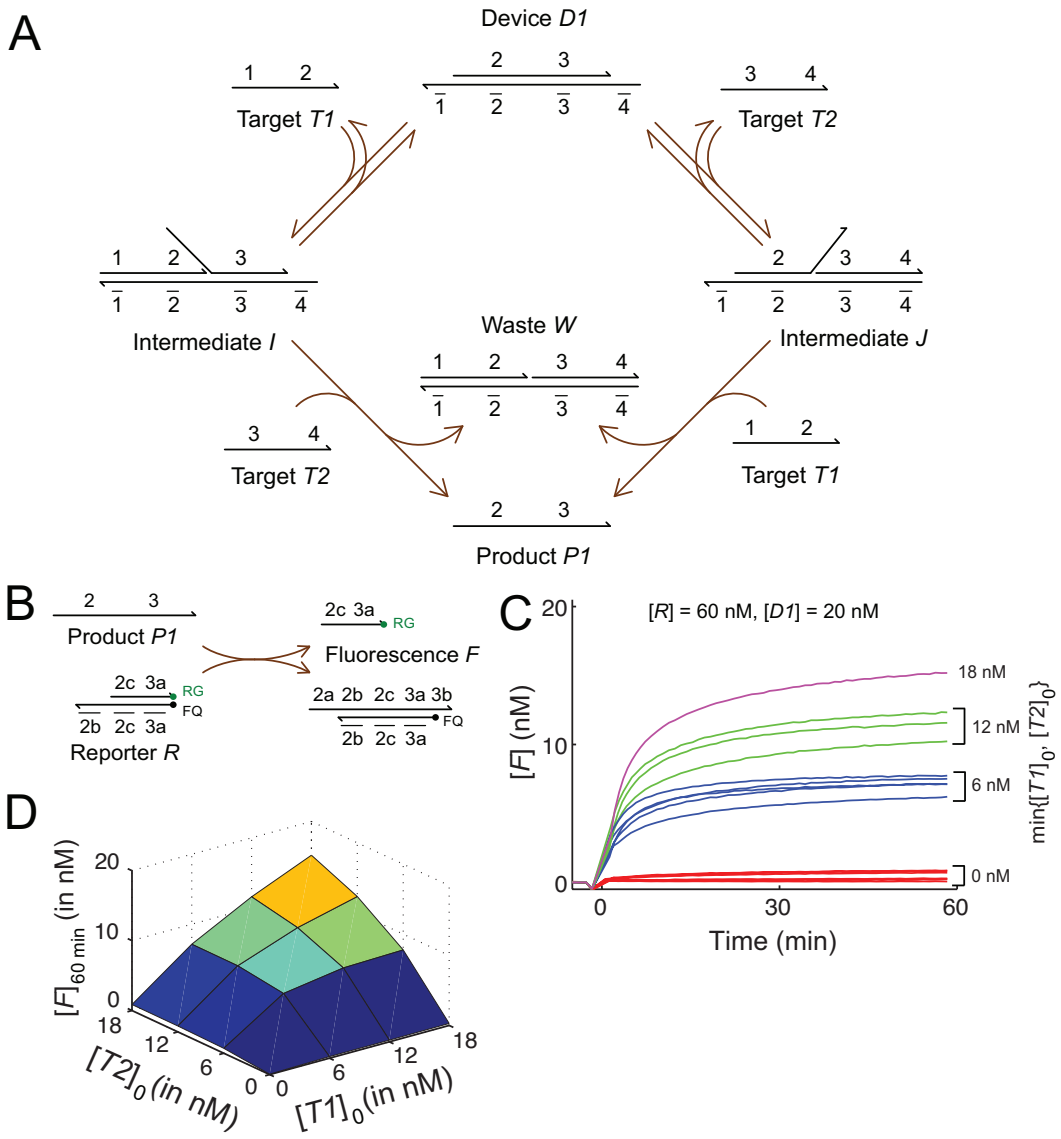


FIG. 7-1: **(A)** Implementation schematic. Numbers represent domains, contiguous bases of DNA that act as a unit in binding. Domain $\bar{x}$ is complementary to domain x, where x is any number. Targets $T1$ and $T2$ must both be present in order for the reaction to proceed to completion and release single-stranded product $P1$. Domain sequences are given in Table 7-1. **(B)** Fluorescent reporter for $P1$. $P1$ reacts stoichiometrically with R and unquenches fluorophore-labelled F. RG denotes the Rhodamine Green fluorophore, and FQ denotes the Iowa Black Fluorescent Quencher. Subdomains 2a, 2b, and 2c comprise domain 2. Subdomains 3a and 3b comprise domain 3. **(C)** Fluorescence results. R and $D1$ were present in solution initially, and appropriate amounts of $T1$ and $T2$ were added at time t = 0 to achieve the final concentrations listed. **(D)** Summary of [F] at t = 1 hr, as a function of target concentrations. The fluorescence at equilibrium should be linear in the minimum of the concentrations of the two targets. The fluorescent value after 1 hour of reaction plotted here is almost linear in the minimum concentration of the two targets; imperfect reaction completion and stoichiometry errors may explain the non-linearity in the data.

reacts stoichiometrically with free $P1$ to release fluorophore-labeled strand F. Fig. 7-1CD shows that the production of $P1$ (and subsequently F) varies directly as the minimum of the concentrations of $T1$ and $T2$. This mechanism is further verified by native polyacrylamide gel electrophoresis (PAGE, see Fig. 7-S1). Domain sequences are given in Table 7-1 (see Text 7-S1 and Table 7-S1 for details).

The stoichiometric and simultaneous consumption of $T1$ and $T2$ allows the concentrations of these species to be compared against one another. In a solution with excess $D1$ at equilibrium, at most one of $T1$ and $T2$ will be present in single-stranded form in significant concentration (the one that was initially in excess). By carefully characterizing the concentration of one standard oligonucleotide, many others can be accurately quantitated by comparison against the standard.

Fig. 7-2A demonstrates a method for concentration comparison based on cooperative hybridization. In this system, the standard oligonucleotide $S1$ is fluorophore-labeled, and its free concentration can be directly assayed. The test solution (boxed) comprises of a quencher-labeled device $D2$ and fluorophore-labeled standard oligonucleotide $S1$ that interact reversibly to form quenched complex L, acting analogously to $T2$ and $D1$ from Fig. 7-1A. At operational conditions, equilibrium favors $D2$ and $S1$, so the test solution exhibits high fluorescence. A sample solution containing target $T3$ is mixed with the test solution. A subsequence of $T3$ (domains 6 and 7) reacts irreversibly with L to form two inert, low-fluorescence waste products $W2$ and $W3$. The equilibrium fluorescence of the mixture decreases linearly with the quantity of $T3$ in the sample solution (see Text 7-S2), subject to a minimum fluorescence when the quantity of target $T3$ exceeds that of the standard $S1$ (Fig. 7-2BC).

The standard $S1$ serves as a subtractive threshold on $T3$ and vice versa; the concentration of the comparison device $D2$ does not significantly factor into the behavior of this reaction network so long as it is in excess. The concentration of $T3$ in the sample solution can be calculated from the equilibrium fluorescence of the solution if $S1$ exceeds $T3$ (see Text 7-S2, Fig. 7-S2, and Table 7-S2). Fig. 7-2D shows that the fluorescence (unquenched $S1$) varies predictably with the concentrations of $S1$ and $T3$.

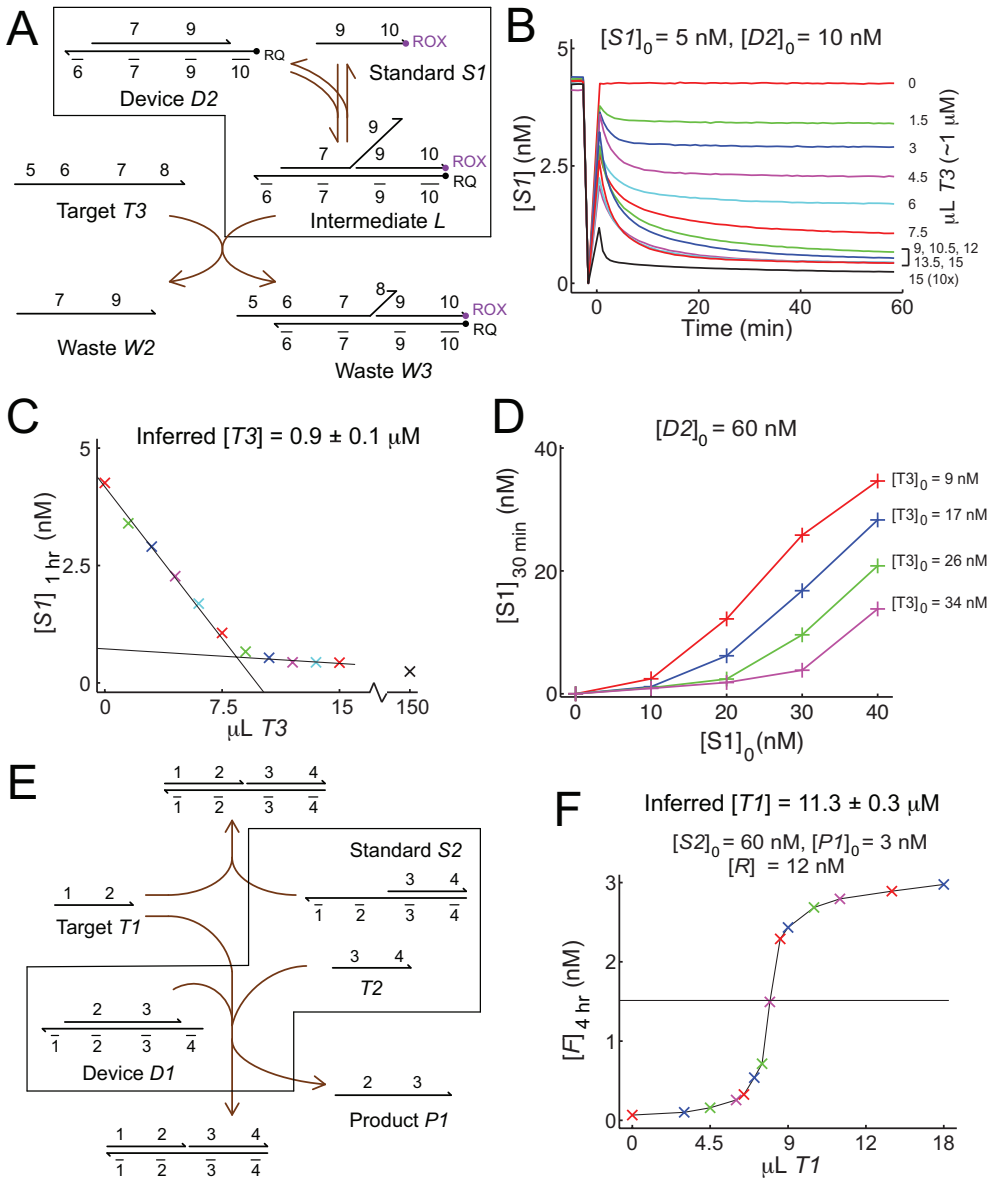


FIG. 7-2: Concentration comparison. **(A)** A subsequence of nucleic acid target ($T3$) cooperatively hybridizes with standard $S1$ to $D2$. Boxed species constitute the test solution; $D2$ exists in excess of $S1$. Fluorescence at equilibrium will be at a minimal baseline level if and only if $[T3]$ exceeds $[S1]$. **(B)** Fluorescence results. Addition of $T3$ at t = 0 causes the fluorescence signal (concentration of free $S1$) to diminish. $[T3]$ was measured by UV absorbance to be 1 μM. "15 (10x)" denotes 15 μL of 10 μM $T3$. **(C)** Fluorescence results from part (B) at t = 1 hr. The fluorescence follows a "kinked line," with the kink at 8.4 ± 1 μL $T3$. Assuming the concentration of $S1$ is correct, the concentration of $T3$ in the sample solution is inferred to be $\frac{1.5 \text{ mL} \cdot 5 \text{ nM}}{8.4 \pm 1 \ \mu\text{L}} = 0.9 \pm 0.1$ μM. **(D)** Tunable thresholding. The concentration of free $S1$ can be tuned by modulating $[T3]$. **(E)** Digital concentration comparison. Standard complex $S2$ is pre-reacted with a small quantity of $P1$ to generate an equal small quantity of $D1$ and $T2$. Target $T1$ preferentially reacts with $S2$, and reacts with $D1$ and $T2$ to release $P1$ only after $S2$ is exhausted. $P1$ reacts with reporter complex R stoichiometrically to release fluorophore-labeled strand F (see Fig. 7-1B). **(F)** Fluorescence signal increases significantly if $T1$ is in excess of $S2$. $[T1]$ was measured by UV absorbance to be 10 μM. Sigmoidal response curve allows precise quantitation. The fluorescence crosses the half-max threshold at 8.0 ± 0.2 μL $T1$, implying $T1$ concentration to be $\frac{1.5 \text{ mL} \cdot 60 \text{ nM}}{8.0 \pm 0.2 \ \mu\text{L}} = 11.3 \pm 0.3$ μM, assuming $T2$ concentration is accurate.

The analog nature of the readout of the design in Fig. 7-2A limits the precision of the concentration comparison and quantitation. A modified design using competitive inhibition (Fig. 7-2E) yields a sigmoidal response curve–a digital answer to whether the target exceeds the standard. In this design, a known amount of standard $S2$ is pre-mixed with a much smaller quantity of $P1$ (0.1x in Fig. 7-2EF) to form an equal quantity of $D1$ and $T2$. After equilibration, reporter R is added, and the resultant solution is the test solution. When a sample solution is added, target $T1$ preferentially reacts with $S2$ to yield an inert waste product. As $S2$ depletes, $T1$ begins reaction with $D1$ and $T2$ to release $P1$, via the cooperative hybridization mechanism. The unreacted $S2$ thus acts as a competitive threshold; when $[T1]$ exceeds $[S2]$, the fluorescence quickly rises to the maximum.

To improve the sensitivity of the concentration comparison devices presented and to restore the concentrations of the excess species, signal amplification (gain) is needed. Here, gain is implemented using a strand displacement-based DNA catalytic reaction, in which a single-stranded DNA molecule (catalyst) catalytically releases single-stranded DNA molecules of independent sequence from multi-stranded complexes. This design is based on Ref. [12], but employs a new set of sequences. Thus, a small quantity of the catalyst DNA, in time, causes the release of a large quantity of product DNA.

Fig. 7-3A shows the modular integration of the concentration comparison network shown in Fig. 7-2A (Box 1) with amplification via DNA catalysis reactions (Box 2) and the fluorescent reporter (Box 3). This synthetic network is capable of amplified digital detection of both over- and under-expression of a target relative to a standard of independent sequence (Fig. 7-3BC). To demonstrate over-expression detection, $T2$ is considered the target of interest and $T1$ serves as the standard; $T1$ and $T2$ annihilate each other via initial reaction, and $T2$ will catalyze the production of CP only when it exceeds the threshold set by $T1$. Likewise for under-expression, $T1$ is the target and $T2$ is the standard, and CP is catalytically produced only when $T1$ is below the threshold set by $T2$.

The kinetics of this system can be modulated by altering the concentrations of CS, CF, and CR. At concentrations of 100 nM, each amplification turnover takes on the order of 2 minutes [13]. For higher amplification turnover or faster kinetics, such as to assay

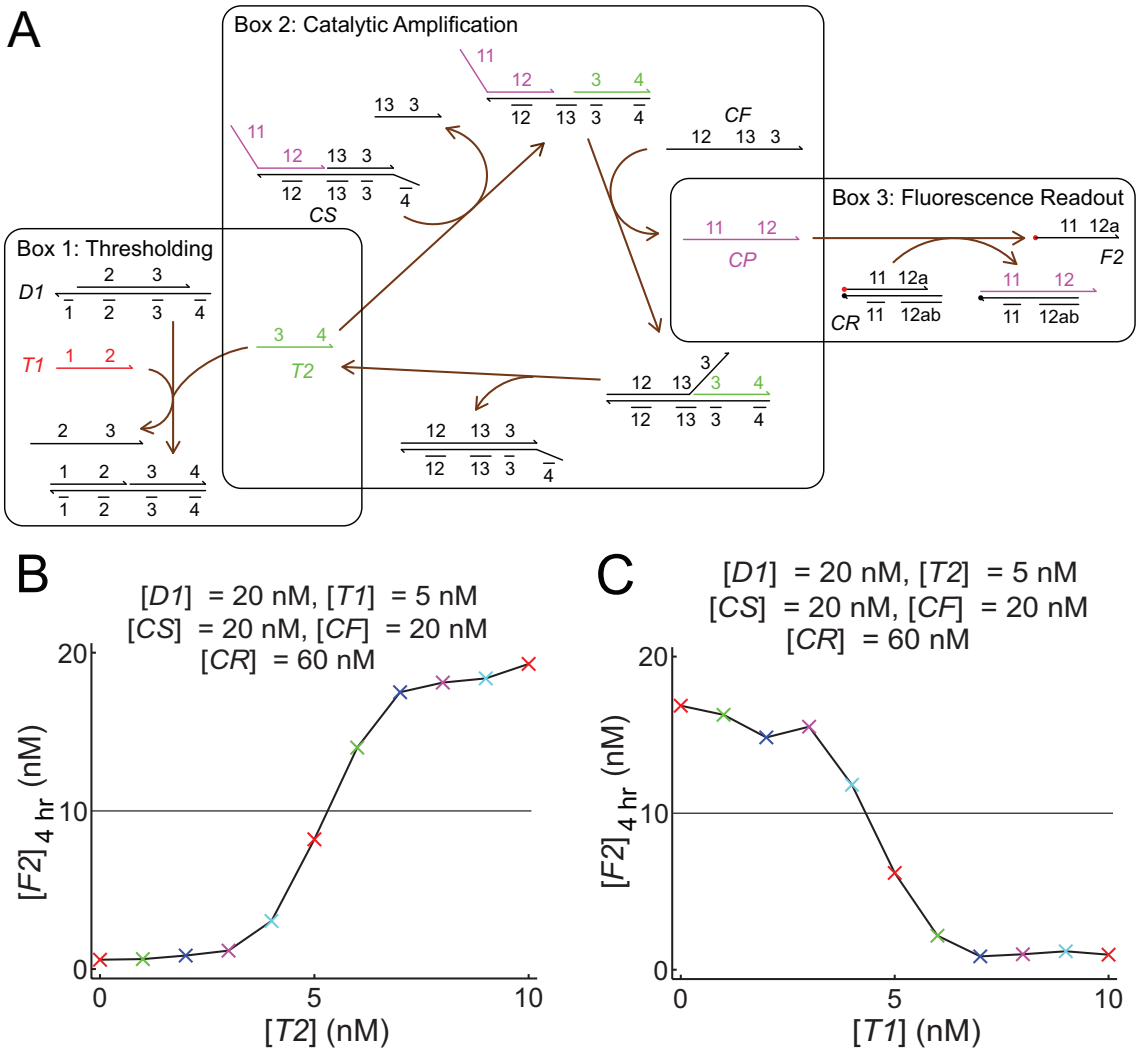


FIG. 7-3: Amplified digital detection of over- and under-expression relative to a threshold. **(A)** Schematic. The detection circuit is divided into 3 modular components, shown as boxes. For over-expression detection, $T1$ serves as the threshold and is present at known concentration. If the amount of $T2$ exceeds that of $T1$, then the excess $T2$ catalyzes the release of CP from CS, which in turn reacts stoichiometrically with fluorescent reporter CR to release fluorophore-labeled strand $F2$. For under-expression detection, $T2$ serves as the threshold with known concentration, and will trigger the catalytic pathway unless excess $T1$ is added. **(B)** Overexpression detection. Fluorescence increases as $[T2]$ exceeds the threshold set by $[T1]$. $T2$ is pre-reacted with $D1$ and $T1$ for 15 min, and then CF, CS, and CR were added to solution to begin the reaction. Maximum fluorescence (gain) is determined by $[CF]$, $[CS]$, and $[CR]$, the concentrations of the catalytic substrate and reporter species. **(C)** Underexpression detection. Fluorescence decreases as $[T1]$ exceeds the threshold set by $[T2]$.

targets with lower concentrations, CP can be amplified further through additional catalysis reactions.

Recently discoveries on microRNAs' roles in gene regulation [14] and their potential as disease markers [15, 16] have motivated the development of high accuracy and high

Dom.	Sequence	Length (nt)
1	5'- CATCACTA -3'	8
2	5'- CTATCATCACACATCTAT -3'	18
3	5'- ACAACCACTTACTTCTTC -3'	18
4	5'- ATCTATCC -3'	8
5	5'- CTATCAT -3'	7
6	5'- CACACAT -3'	7
7	5'- CTATACAACCACTTACTT -3'	18
8	5'- CTTC -3'	4
9	5'- GCCATCAGAACTTAACCT -3'	18
10	5'- AACTC -3'	5

Table 7-1: Domain sequences (see also Table 7-S1).

sensitivity methods for quantitating short oligonucleotides in native biological conditions. Digital concentration comparison methods presented here can be used to perform accurate nucleic acid quantitation. Although the sensitivity of the methods developed here are significantly lower than that of qPCR, enzyme dependence renders the latter difficult to use for *in situ* and *in vivo* applications. Molecular beacons have been used for *in situ* imaging and gene profiling [17], but lack gain and are limited in sensitivity.

The reactions shown here were rationally designed based on the biophysics of nucleic acid hybridization, branch migration, and dissociation. Previous characterization of similar reactions have demonstrated that they function robustly across a wide range of solution salinities and temperatures [12], as well as in the presence of total RNA and cell lysate [5, 12]. Furthermore, similar strand displacement-based RNA devices have been made to assay and regulate gene expression in cells [18, 19]. Thus, it is possible that many of these devices can be made to function *in situ* and *in vivo* for assaying and regulating gene expression.

The methods presented in this paper form a general framework for amplified concentration comparison, which could potentially be extended to non-nucleic acid targets via integration with aptamers [21, 22]. Previous works have demonstrated that aptamers can be rationally designed to influence hybridization thermodynamics [23–25], which can in turn be used to trigger hybridization, strand displacement, and dissociation reactions [24, 25].

One other potential application for oligonucleotide concentration comparison and quantitation is point-of-care diagnosis. In order realize "instrument-free" detection, a visual assay is required. This could potentially be implemented by integrating the methods presented here with the DNA-induced gold colloid aggregation [26] (causing a visible pink to purple

color change) or with Guanine-quartet peroxidase activity [27] (causing visible bubbling).

The cooperative hybridization mechanism enables more than just nucleic acid detection and quantitation; it is a fundamentally new design primitive that allows simultaneity detection, precise timing control, and non-linear signal responses. Cooperative hybridization can be used to engineer a variety of nucleic acid reactions and networks, such as logical networks and linear classifiers, and will be instrumental in the construction of complex synthetic reaction networks for controlling biochemistry.

The complexity of natural biochemical circuits enables wondrous behaviors of life such as development, metabolism, and reproduction. It remains an outstanding goal of synthetic biology to rationally design reaction networks that exhibit similar spatial/temporal control of biochemistry; the design, demonstration, and integration of modular nucleic acid systems is one promising approach. The concentration sensing systems demonstrated in this paper could be an important component of nucleic acid logical and control networks.

Methods

Buffer conditions. DNA oligonucleotides were resuspended and stored in TE buffer (10 mM Tris · HCl pH balanced to 8.0, with 1 mM EDTA·Na_2, purchased as 100x stock from Sigma-Aldrich) at 4 °C. Directly preceding experiments, TE buffer with 62.5 mM $MgCl_2$ was added at 1:4 ratio to the sample, achieving a final $MgCl_2$ concentration of 12.5 mM. Because roughly 1 mM of the Mg^{2+} is chelated by the EDTA present in solution, the free concentration of Mg^{2+} is estimated to be 11.5 mM. All experiments and purifications were performed at 25 °C.

DNA oligonucleotides. DNA oligonucleotides used in this study were purchased from Integrated DNA Technologies (IDT). Where applicable, fluorophores were attached by IDT as well. $T3$, $S1$, and the oligonucleotides that comprise complexes R, $D2$, and CR were purified by IDT via high performance liquid chromotography (HPLC).

Multistranded complexes $D1$ ad $S2$ were purified by hand via native PAGE as follows: Oligonucleotides needed for each sample were prepared with nominally correct stoichiometry at 20 μM and annealed with an Eppendorf Mastercycler Gradient thermocycler. The

solutions were brought down from 95 °C to 20 °C at a constant rate over the course of 90 minutes. The solution were then run on 12% native PAGE at 180 V for 6 hours. The proper bands were cut out and eluted in 1 mL TE/Mg^{2+} buffer for 2 days. Typical yields ranged from 40% to 60%. The concentrations of purified complexes were estimated by measuring absorbance at 260 nm using an Eppendorf Biophotometer, and dividing by extinction coefficients for single- and double-stranded DNA predicted by nearest-neighbor models.

Spectrofluorimetry studies. Spectrofluorimetry studies were done using a SPEX Fluorolog-3 (Horiba) with 1.6 mL synthetic quartz cells (Hellma 119-004F). Experiments were done with integration time of 10 seconds for every 60 second time-point.

F was labeled with the Rhodamine Green (RG) fluorophore. For experiments measuring the concentration of F, the excitation wavelength was set to 510 nm, while emission wavelength was set to 531 nm, with 2 nm band pass slits for both monochromators.

$S1$ was labeled with the carboxy-X-rhodamine (ROX) fluorophore. For experiments measuring the concentration of $S1$, the excitation wavelength was set to 588 nm, while emission wavelength was set to 602 nm, with 4 nm band pass slits for both monochromators.

Prior to each experiment, all cuvettes were cleaned thoroughly: each cuvette was washed 15 times in distilled water, once in 70% ethanol, another 5 times in distilled water, and finally once more in 70% ethanol.

For the slit size, concentrations, and times chosen, no measurable photobleaching was observed. Fluorescence measurements are linear in the concentration of the free fluorescent oligonucleotides (F or $S1$). All experimental results were within the linear regime of the spectrofluorimeter detector, according to specification sheets provided by the manufacturer.

Fluorescence normalization. Fluorescence is normalized so that 1 normalized unit (n.u.) of fluorescence corresponds to 1 nM of unquenched fluorophore-labeled strand F or $S1$. Day-to-day and sample-to-sample variations are estimated to be less than 5% [12].

Acknowledgements. The author thanks Tosan Omabegho, Paul W. K. Rothemund, and Georg Seelig for useful suggestions regarding the organization and presentation of the manuscript. The author is particular grateful to Erik Winfree for financial support for ex-

perimental reagents. This work was supported in part by NSF grants 0832824 and 0728703. DYZ is supported by the Fannie and John Hertz Foundation.

There is a patent pending on the methods demonstrated in this paper.

7-S1. DNA sequence design

The lengths and sequences of the 2 and 3 domains were chosen so that spontaneous dissociation of $P1$ from state I or J would be extremely unlikely. For average 15-mers (with $\Delta G^0 \approx 21$ kcal/mol), the dissociation time is estimated to be roughly 10 years, which is far longer than the timescale of the reactions being monitored. However, in order to account for synthesis errors that could destabilize the binding of these domains to their complements, the lengths of the 2 and 3 domains were chosen to be 18 nt. It is expected that the devices would work similarly for domains 2 and 3 of equal or longer length, except insofar as the kinetics of branch migration would be slowed for long domains.

There are two criteria on the lengths and sequences of the 1 and 4 domains:

First, the thermodynamics of 1 and 4 binding to their complement domains in $D1$ should overcome the configuration entropy loss of colocalizing another molecule ($T1 + T2 + D1 \rightarrow P1 + W$). That is, $\Delta G^\circ(1) + \Delta G^\circ(4) < \Delta G^\circ_{init} + RT\ln(c/1\text{ M})$, where $\Delta G^\circ(1)$ and $\Delta G^\circ(4)$ represent the standard free energies of binding of the 1 and 4 domains to their complement, and ΔG°_{init} represents the free energy of duplex initiation (approximated to be -1.8 kcal / mol at 25 °C [1]), and c represents the concentration of $D1$. For 10 nM concentrations, $\Delta G^\circ(1) + \Delta G^\circ(4) < -12.8$ kcal / mol. Assuming $\Delta G^\circ(1) \approx \Delta G^\circ(4)$, $\Delta G^\circ(1) < -6.4$ kcal/mol. This value roughly corresponds to lower bound of approximately 5 nt for domains with roughly equal distribution of G/C and A/T bases.

Second, the thermodynamics of 1 and 4 binding to their complement domains in $D1$ should be weak enough so that if only one target is present, then that target is thermodynamically favored to be not bound to $D1$. The upper limit to the binding strengths of the 1 and 2 domains therefore depend on the operational concentrations:

$\Delta G^\circ(1), \Delta G^\circ(4) \geq RT\ln(c) + \Delta G^\circ_{init} = -12.8$ kcal / mol, corresponding to roughly 10 nt for domains with roughly equal distribution of G/C and A/T bases. Consequently, the lengths of both the 1 and 4 domains were chosen to be 8 nt.

The domain sequences used in this paper were designed to be minimally interacting. Special emphasis was placed on ensuring that single-stranded targets and products ($T1$, $T2$, $P1$, $S1$) possessed no significant secondary structure. Although it is expected that the designs will function qualitatively similarly for strands with minor secondary structure, secondary structure is known to slow down the kinetics of hybridization and branch migration processes [30].

Dom.	Sequence	Length (nt)
1	5'- CATCACTA -3'	8
2 = 2a:2b:2c	5'- CTATCATCACACATCTAT -3'	18
2a	5'- CTATCAT -3'	7
2b	5'- CACACAT -3'	7
2c	5'- CTAT -3'	4
3 = 3a:3b	5'- ACAACCACTTACTTCTTC -3'	18
3a	5'- ACAACCACTTACTT - 3'	14
3b	3'- CTTC -3'	4
4	5'- ATCTATCC -3'	8
5	5'- CTATCAT -3'	7
6	5'- CACACAT -3'	7
7	5'- CTATACAACCACTTACTT -3'	18
8	5'- CTTC -3'	4
9	5'- GCCATCAGAACTTAACCT -3'	18
10	5'- AACTC -3'	5
11	5'- CTTTCCTACA -3'	10
12	5'- CCTACGTCTCCAACTAACTTACGG -3'	24
12a	5'- CCTACGTCTC -3'	10
12b	5'- CAACTAA -3'	7
12c	5'- CTTACGG -3'	7
13	5'- CCCTC -3'	5

Table 7-S1: Domain sequences.

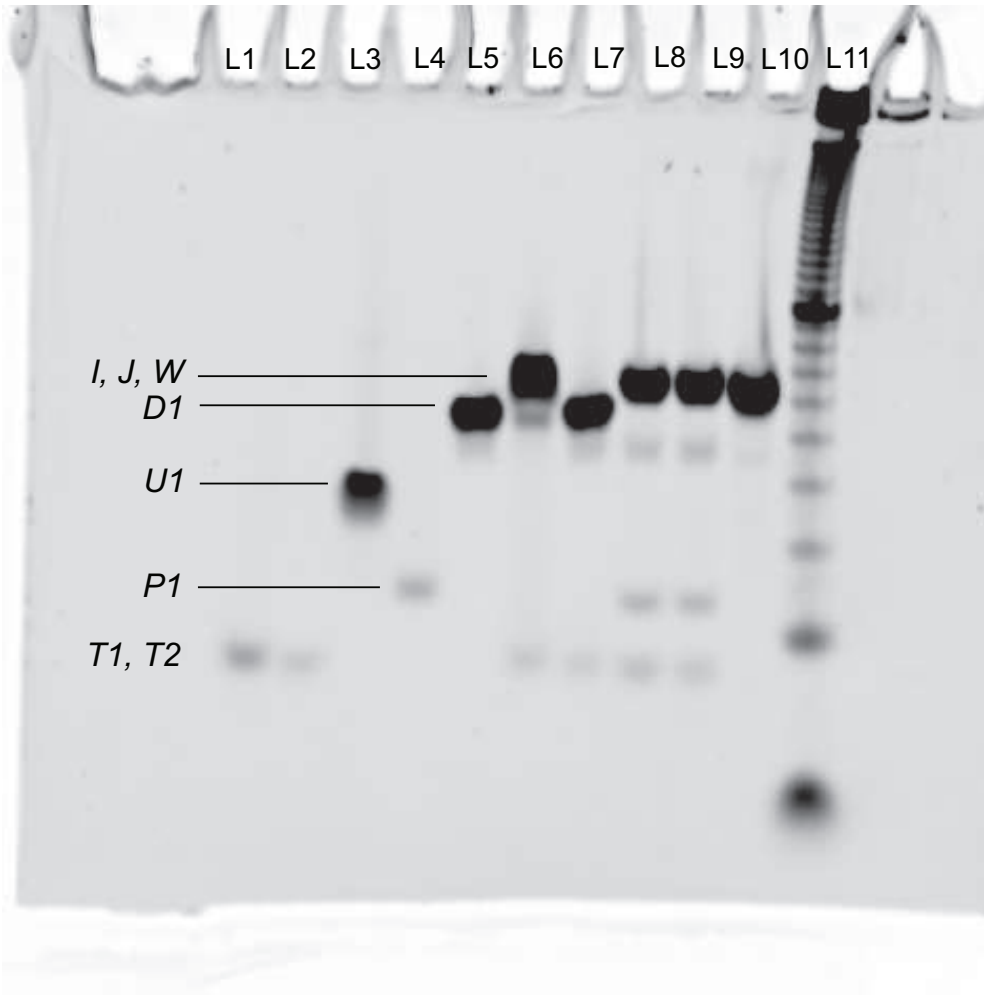


FIG. 7-S1: Native polyacrylamide gel electrophoresis (PAGE) on cooperative strand displacement. This experiment used 12% acrylamide (19:1 acrylamide:bis), diluted from 40% acrylamide stock (Ambion). Native loading dye containing xylene cyanol-FF (XCFF) in 50% glycerol was added to all samples, achieving final glycerol concentration of 10% by volume. Gels were run at 25 °C using a Novex chamber with external temperature bath. Gels were stained with Sybr-Gold stain (Invitrogen), and scanned with a Bio-Rad Molecular Imager. "30 min" denotes that the strands and complexes were mixed and allowed to react at 25 °C for 30 minutes. $U1$ is the name of the bottom strand of $D1$ ($U1$ and $P1$ together comprise $D1$). $T1$, $T2$, and $P1$ stains less efficiently than other strands and complexes due to their short lengths and single-stranded nature. The single-stranded $P1$ band appears only when both $T1$ and $T2$ are present. Intermediate I migrates at approximately the same speed as W. Note that in lane 6, I does not significantly dissociate on the time scale of the gel. This is postulated to be because of a combination higher operational concentrations (200 nM for gel vs. 20 nM for fluorescence experiments) and strand impurity. Intermediate J is assumed to dissociate on the time scale of running the gel.

Lane	Contents	Lane	Contents
1	$T1$ (200 nM)	7	$D1$ (200 nM) + $T2$ (400 nM) [30 min]
2	$T2$ (200 nM)	8	$D1$ (200 nM) + $T1$ (400 nM) + $T2$ (400 nM) [30 min]
3	$U1$ (200 nM)	9	$D1$ (200 nM) + $T1$ (400 nM) + $T2$ (400 nM) [annealed]
4	$P1$ (200 nM)	10	W (200 nM) [annealed]
5	$D1$ (200 nM)	11	10 nt duplex ladder
6	$D1$ (200 nM) + $T1$ (400 nM) [30 min]		

7-S2. Concentration calculation from fluorescence.

There are three sources of fluorescence signal in the experiments described in Fig. 7-2: the unquenched fluorophores in $S1$, the quenched fluorophores in L, and the background fluorescence. Define b to be the background fluorescence. Define a to be the fluorescence per unit of the unquenched fluorophore. Define q to be the quenching ratio of the fluorophore (the fluorescence of the unquenched fluorophore divided by the fluorescence of the quenched fluorophore). Thus, the fluorescence f of a solution can be written as:

$$\begin{aligned} f &= a[S1] + \frac{a}{q}[L] + b \\ &= \frac{a(q-1)}{q}[S1] + \frac{a[S1]_{\text{tot}}}{q} + b \end{aligned}$$

where $[S1]_{\text{tot}} = [S1] + [L]$ is the total concentration of $S1$ in solution. As can be seen, the fluorescence f is linear in $[S1]$.

Idealized case.

In the idealized case of when the sample solution containing the target reaches equilibrium with the test solution containing the standard and device, and where the fluorescence readout is infinitely accurate, the concentration of the target in the sample solution can be easily calculated from a single experiment.

Define x to be the volume of sample solution that reacted with y volume of test solution. The system operates under the assumption that $y \geq x$. Define the concentration of $T3$ in the sample solution to have been $[T3]_s$. Define the total concentrations of $S1$ and $D2$ (including the portion sequestered in L) in the test solution to have been $[S1]_t$ and $[D2]_t$, respectively. Thus, there are $x[T3]_s$ moles of $T3$, $y[S1]_t$ moles of $S1$, and $y[D2]_t$ moles of $D2$. In order for quantitation to be possible, $y[S1]_t > x[T3]_s$ (i.e. standard exceeds target).

Define $[S1]_m$, $[D2]_m$, and $[L]_m$ to be the concentrations of free $S1$, free $D2$, and intermediate L in the mixture of the sample and test solutions at equilibrium. $[S1]_m$ is inferred

from the fluorescence level, and we wish to derive $[T3]_s$, the concentration of the target in the initial sample solution.

$$\begin{aligned}\frac{y[S1]_t - x[T3]_s}{x+y} &= [S1]_m + [L]_m \\ \frac{y[D2]_t - x[T3]_s}{x+y} &= [D2]_m + [L]_m\end{aligned}$$

Reorganizing,

$$\begin{aligned}[L]_m &= \frac{y[S1]_t - x[T3]_s}{x+y} - [S1]_m \\ [D2]_m &= [S1]_m + \frac{y([D2]_t - [S1]_t)}{x+y}\end{aligned}$$

From the equilibrium of the mixture:

$$\begin{aligned}K_{eq} &= \frac{[L]_m}{[S1]_m[D2]_m} \\ &= \frac{\frac{y[S1]_t - x[T3]_s}{x+y} - [S1]_m}{[S1]_m([S1]_m + \frac{y([D2]_t - [S1]_t)}{x+y})}\end{aligned}$$

$$K_{eq}[S1]_m([S1]_m + \frac{y([D2]_t - [S1]_t)}{x+y}) + [S1]_m = \frac{y[S1]_t - x[T3]_s}{x+y}$$

$$[T3]_s = \frac{y[S1]_t}{x} - \frac{K_{eq}y([D2]_t - [S1]_t) + x + y}{x}[S1]_m - \frac{K_{eq}(x+y)}{x}[S1]_m^2 \qquad (1)$$

Every variable on the right hand side of the equation is known (with K_{eq} calculated from the ΔG° of the reaction); thus, $[T3]_s$ can be calculated from observed fluorescence.

Quantitation using a calibration series.

Given that at experimental concentrations, the reactions could take quite long to reach equilibrium, it is possible instead to use a series of experiments with various different volumes of the sample solution reacting with the test solution. This is the method used to quantitate the concentration of $T3$ shown in Fig. 7-2BC. As the volume of the sample solution increases, the amount of $T3$ eventually overtakes that of the standard $S1$. The volume of sample solution at which the amount of $T3$ stoichiometrically matches that of $S1$ is here denoted as the *matching volume*.

When the matching volume of the sample solution is added to the test solution, then at equilibrium all of the $S1$ will be sequestered in waste $W3$ and quenched. Further addition of sample solution will only slightly decrease the fluorescence (due to dilution of the quenched fluorophore in $W3$). It takes a long time for equilibrium to be established when the matching volume of the sample solution is added due to the second-order nature of the $L + T3 \rightarrow W2 + W3$ reaction, but equilibrium is more quickly approached when the amount of sample solution added deviates significantly from the matching volume (the reaction kinetics become pseudo-first order due to the excess of either L or $T3$).

Given the fluorescence of various mixtures with different amounts of sample solution at a particular time point, the matching volume can be estimated by the intersection of two linear fits of fluorescence to sample solution volume (Fig.7-S2). See the following section "Pseudo-linearity" for the justification on why fluorescence is expected to linearly decrease with the amount of $T3$ added. The data from the later time points are expected to be more reliable because the reaction has proceeded closer to equilibrium, and because the earlier data points may be affected more strongly by incomplete mixing and experimental error on the reported time points. The linear fits in Fig. 7-S2 use 4 data points each; Table 7-S2 show the inferred matching volumes based on linear fits using 3, 4, and 5 data points. As can be seen, the inferred matching volumes are quite similar regardless of how many data points are used in the linear fits.

From the data shown, the matching volume is seen to be 8.4 ± 1.0 μL. Assuming that the concentration of the standard $S1$ is accurate at 5 nM, the total amount of $S1$ is 5 nM $\cdot$1.5 mL= 7.5 pmol, and the inferred concentration of $T3$ in the sample solution

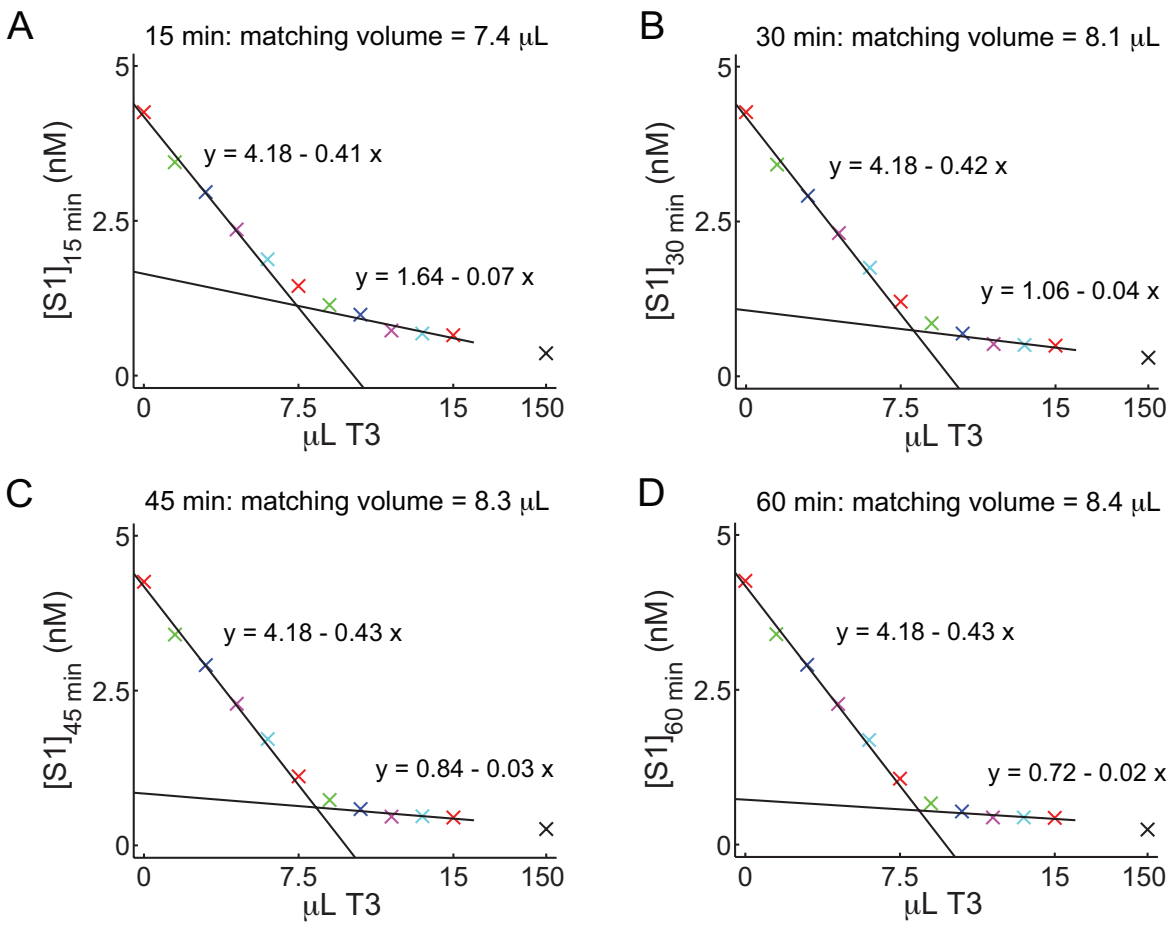


FIG. 7-S2: Concentration inference from the data in Fig. 7-2B. The fluorescence levels of the various traces at 15 min, 30 min, 45 min, and 60 min are plotted in **(A)**, **(B)**, **(C)**, and **(D)**, respectively. Each sub-figure shows two linear fits, one to the first 4 data points and one to the last 4 data points (other than the 150 μL). The intersection of the two linear fits is the inferred matching volume of $T3$, the volume of $T3$ in which the quantity of $T3$ and $S1$ are stoichiometrically balanced.

is is $\frac{7.5\text{pmol}}{8.4\pm1.0\mu\text{L}} = 0.89 \pm 0.12$ μM, or 0.9 ± 0.1 μM after accounting for significant digits.

	3 points	4 points	5 points
15 min	7.8 μL	7.4 μL	7.5 μL
30 min	8.1 μL	8.1 μL	8.1 μL
45 min	8.3 μL	8.3 μL	8.3 μL
60 min	8.3 μL	8.4 μL	8.4 μL

Table 7-S2: Inferred matching volume of $T3$ from data in Fig. 7-2B.

Pseudo-linearity.

We wish to show that the fluorescence ($[S1]_m$) decreases pseudo-linearly with the amount

of $T3$ added ($x[T3]_s$). Rearranging eq. (5),

$$K_{eq}(x+y)[S1]_m^2 + (K_{eq}y([D2]_t - [S1]_t) + x + y)[S1]_m - (y[S1]_t - x[T3]_s) = 0$$

Solving this quadratic relation,

$$\begin{aligned}[S1]_m &= -\frac{K_{eq}y([D2]_t - [S1]_t) + x + y}{2K_{eq}(x+y)} \\ &\quad + \frac{\sqrt{(K_{eq}y([D2]_t - [S1]_t) + x + y)^2 + 4K_{eq}(x+y)(y[S1]_t - x[T3]_s)}}{2K_{eq}(x+y)}\end{aligned}$$

The second term can be approximated as

$$\frac{1}{2K_{eq}(x+y)}(x + y + K_{eq}y([D2]_t - [S1]_t) + 2K_{eq}(y[S1]_t - x[T3]_s))$$

if $2K_{eq}(y[S1]_t - x[T3]_s)$ is small compared to $x + y + K_{eq}y([D2]_t - [S1]_t)$. We will show that this is the case shortly. For now, we continue using the approximation, substituting back into the expression for $[S1]_m$:

$$\begin{aligned}[S1]_m &\approx -\frac{K_{eq}y([D2]_t - [S1]_t) + x + y}{2K_{eq}(x+y)} + \frac{x + y + K_{eq}y([D2]_t - [S1]_t) + 2K_{eq}(y[S1]_t - x[T3]_s)}{2K_{eq}(x+y)} \\ &= \frac{2K_{eq}(y[S1]_t - x[T3]_s)}{2K_{eq}(x+y)} \\ &= \frac{y[S1]_t}{x+y} - \frac{1}{x+y}x[T3]_s\end{aligned}$$

The final relationship between $[S1]_m$ and $x[T3]_s$ is thus seen to be approximately linear. This pseudo-linearity depends upon the condition:

$$2K_{eq}(y[S1]_t - x[T3]_s) \ll x + y + K_{eq}y([D2]_t - [S1]_t)$$

For our purpose, let us solve for the conditions under which:

$$2K_{eq}(y[S1]_t - x[T3]_s) < 0.5(x + y + K_{eq}y([D2]_t - [S1]_t))$$

When this is true, the difference between the approximate $[S1]_m$ and the real $[S1]_m$ is less than 20%. Rearranging this relation,

$$4K_{eq}(y[S1]_t - x[T3]_s) \ < \ x + y + K_{eq}y([D2]_t - [S1]_t)$$

The left hand side is smaller than $4K_{eq}y[S1]_t$, and the right hand side is greater than y, so a sufficient condition is:

$$\begin{aligned} 4K_{eq}y[S1]_t \ &< \ y \\ K_{eq} \ &< \ \frac{1}{4[S1]_t} \end{aligned}$$

Recall that $K_{eq} = \frac{[L]}{[S1][D2]}$.

$$\begin{aligned} \frac{[L]}{[S1][D2]} \ &< \ \frac{1}{4[S1]_t} \\ \frac{[L]}{[S1]} \ &< \ \frac{[D2]}{4[S1]_t} \end{aligned}$$

When $[D2]_t \geq 2[S1]_t$, $[D2]$ is necessarily greater than $[S1]_t$, so the right hand side is

greater than $\frac{1}{4}$. Thus, the necessary condition for the approximation is satisfied whenever $\frac{[L]}{[S1]} < \frac{1}{4}$. This equates to saying that operational concentrations must be low enough that the $S1 + D2 \rightleftharpoons L$ reaction favors the reactants by at least a 4:1 ratio.

[1] SantaLucia, J. & Hicks, D. The thermodynamics of DNA structural motifs. *Ann. Rev. Biochem.* **33,** 415-440 (2004).

[2] Zhang, D. Y. & Winfree, E. Control of DNA Strand Displacement Kinetics Using Toehold Exchange. *J. Am. Chem. Soc.* **131,** 17303-17314 (2009).

[3] Green, S., Bath, J. & Turberfield, A. J. Coordinated chemomechanical cycles: a mechanism for autonomous molecular motion. *Phys. Rev. Lett.* **101,** 238101 (2008).

[4] Omabegho, T., Sha, R. & Seeman, N. C. A bipedal DNA Brownian motor with coordinated legs. *Science* **324,** 67-71 (2009).

[5] Seelig, G., Soloveichik, D., Zhang, D. Y. & Winfree, E. Enzyme-free nucleic acid logic circuits. *Science* **314,** 1585-1588 (2006).

[6] Ran, T., Kaplan, S. & Shapiro, E. Molecular implementation of simple logic programs. *Nat. Nanotech.* **4,** 642-648 (2009).

[7] Venkataraman, S., Dirks, R. M., Rothemund, P. W. K., Winfree, E. & Pierce, N. An autonomous polymerization motor powered by DNA hybridization. *Nat. Nanotech.* **2,** 490-494 (2007).

[8] Yin, P., Choi, H. M. T., Calvert, C. R. & Pierce, N. A. Programming biomolecular self-assembly pathways. *Nature* **451,** 318-322 (2008).

[9] Goodman, R. P. *et al.* Reconfigurable, braced, three-dimensional DNA nanostructures. *Nat. Nanotech.* **3,** 93-96 (2008).

[10] Andersen, E. S. *et al.* Self-assembly of a nanoscale DNA box with a controllable lid. *Nature* **459,** 73-76 (2009).

[11] Soloveichik, D., Seelig, G. & Winfree, E. DNA as a universal substrate for chemical kinetics. *Proc. Nat. Acad. Sci.* (pre-published online, 2010. DOI: 10.1073/pnas.0909380107).

[12] Zhang, D. Y., Turberfield, A. J., Yurke, B. & Winfree, E. Engineering entropy-driven reactions and networks catalyzed by DNA. *Science* **318,** 1121-1125 (2007).

[13] Zhang, D. Y. & Winfree, E. Robustness and modularity properties of a non-covalent DNA catalytic reaction. *Nuc. Acid Res.* (2010, pre-published online DOI:10.1093/nar/gkq088).

[14] Bartel, D. P. MicroRNAs: target recognition and regulatory functions. *Cell* **136,** 215-233 (2009).

[15] Lu, J. *et al.* MicroRNA expression profiles classify human cancers. *Nature* **435,** 834-838 (2005).

[16] van Rooij, E. *et al.* A signature pattern of stress-responsive microRNAs that can evoke cardiac hypertrophy and heart failure. *Proc. Nat. Acad. Sci.* **103,** 18255-18260 (2005).

[17] Peng, X. H. *et al.* Real-time detection of gene expression in cancer cells using molecular beacon imaging: new strategies for cancer research. *Cancer Res.* **65,** 1909-1917 (2005).

[18] Seferos, D. S., Giljohann, D. A., Hill, H. D., Prigodich, A. E. & Mirkin, C. A. Nano-flares: probes for transfection and mRNA detection in living cells. *J. Am. Chem. Soc.* **129,** 15477-15479 (2007).

[19] Isaacs, F. J. *et al.* Engineered riboregulators enable post-transcriptional control of gene expression. *Nat. Biotechnol.* **22,** 841-847 (2004).

[20] Tyagi, S. & Kramer, F. R. Molecular beacons: probes that fluoresce upon hybridization. *Nat. Biotech.* **14,** 303-308 (1996).

[21] Ellington, A. D. & Szostak, J. W. In vitro selection of RNA molecules that bind specific ligands. *Nature* **346,** 818-822 (1990).

[22] Li, N. *et al.* Aptamers that recognize drug-resistant HIV-1 reverse transcriptase. *Nuc. Acids Res.* **36,** 6739-6751 (2008).

[23] Penchovsky R. & Breaker R. R. Computational design and experimental validation of oligonucleotide-sensing allosteric ribozymes. *Nature Biotech.* **23,** 1424-1433 (2005).

[24] Dirks, R. M. & Pierce, N. A. Triggered Amplification by Hybridization Chain Reaction. *Proc. Nat. Acad. Sci.* **101,** 15275-15278 (2004).

[25] Zheng, D., Seferos, D. S., Giljohann, D. A., Patel, P. C. & Mirkin, C. A. Aptamer nano-flares for molecular detection in living cells. *Nano Lett.* **9,** 3258-3261 (2009).

[26] Elghanian, R., Storhoff, J. J., Mucic, R. C., Letsinger, R. L. & Mirkin, C. A. Selective colorimetric detection of polynucleotides based on the distance-dependent optical properties of gold nanoparticles. *Science,* **277,** 1078-1081 (1997).

[27] Cheng, X., Liu, X., Bing, T., Cao, Z. & Shangguan, D. General peroxidase activity of G-quadruplex-hemin complexes and its application in ligand screening. *Biochemistry* **48,** 7817-7823 (2009).

[28] Puglisi, J. D. & Tinoco, I. Absorbance melting curves of RNA. *Methods in Enzymology* **180,** 304-325 (1989).

[29] Owczarzy, R., Moreira, B. G., You, Y., Behlke, M. A. & Walder, J. A. Predicting stability of DNA duplexes in solutions containing Magnesium and monovalent cations. *Biochemistry* **47,** 5336-5353 (2008).

[30] Gao, Y., Wolf, L. K.& Georgiadis, R. M. Secondary structure effects on DNA hybridization kinetics: a solution versus surface comparison. *Nucleic Acids Research* **34,** 3370-3377 (2006).

Chapter 8: Characterizing Cooperative Hybridization

Richard Feynman is often quoted for saying "What I cannot create, I do not understand." I would argue that Feynman's condition for understanding was necessary but not sufficient: a baker could easily create a tasty cake by meticulously following a well-written set of instructions, yet have no understanding of the chemical reactions that occurred to make the finished cake fluffy and moist. Rote memory of the recipe is insufficient for the baker to dynamically adjust his recipe to suit different tastes.

Similarly in nucleic acid engineering, true understanding requires more than an experimental procedure that yields consistent desireable results. Careful modeling of the intermediate reactions that occur within individual components is necessary for a deep level of comprehension of the relevant biophysics, which is in turn necessary for reliable and modular integration of many components and reactions. For this reason, this chapter focuses on characterizing the cooperative hybridization primitive introduced in the previous chapter.

This work builds reaction models similar to those shown in Chapter 3, and similarly measures the rate constants of intermediates reactions. The agreement between the model and the experimental data is generally good, although kinetics and thermodynamics are significantly less predictable when fluorophores and/or quenchers are functionalized directly to the hybridization target.

In addition to characterizing cooperative hybridization, this work also explores the use of cooperative hybridization as potential components of analog and digital nucleic acid logic circuits. Cooperative hybridization-based implementations of logical AND and NOT gates are presented and experimentally demonstrated. The AND and NOT gates were also cascaded to form a circuit that exhibits NAND logic. Compared to previous implementations of enzyme-free nucleic acid logic gates and circuits, cooperative hybridization-based implementations excel in being robust to impurities and easily thresholdable.

This project splintered off from the project described in Chapter 7 during 2010, when I realized that I was not going to be able to tell a coherent and concise story about both the characterization and the biotechnological applications of cooperative hybridization in a single 5-page manuscript.

The version presented in this chapter is not a final manuscript. This work is in preparation for submission as:

Zhang, David Yu. "Cooperative Hybridization as a Primitive for Engineering Nucleic Acid Circuits."

Nucleic acids have been demonstrated to be versatile nanoscale engineering materials with the construction of dynamic DNA structures, motors, and circuits. These constructions generally rely on the clever use and integration of relatively few reaction mechanisms and design primitives. Cooperative hybridization is a recently introduced reaction mechanism in which two oligonucleotides stoichiometrically, simultaneously, and cooperatively hybridize to a DNA complex. Here, cooperative hybridization is rigorously characterized and modeled. Additionally, cooperative hybridization is shown to implement digital AND, NOT, and NAND logic, as well as analog minimum and subtraction functions. Compared to previous implementations of nucleic acid logic components and circuits, designs based on cooperative hybridization excel in being robust to impurities and not requiring purification.

Introduction

Precise understanding and control of nucleic acid thermodynamics [1, 2] and kinetics [3, 4] have enabled the construction of nanoscale nucleic acid structures [5–8, 10–12], motors [13–16], and circuits [17, 18]. The design of many of these impressive nucleic acid system are based on the clever and repeated use of simple but reliable mechanisms, such as toehold-mediated strand displacement [13, 16–18], self-assembly of rigid DNA monomers [5, 9, 10], and DNA origami [7, 11, 12].

Dynamic DNA nanotechnology, in which the focus is on the non-equilibrium behavior of designed nucleic acid systems, has primarly relied on the first primitive. In toehold-mediated strand displacement, a single-stranded nucleic acid molecule (henceforth *strand*) binds via its toehold domain to a partially double-stranded complex, and releases one or more product strands. Toehold-mediated strand displacement is a powerful primitive that has enabled the design of logical gates and circuits [17, 19], catalytic systems [18, 20, 21], and allosteric nucleic acids [22].

Despite its many functionalities, the toehold-mediated strand displacement mechanism has not been experimentally shown to perform simultaneity detection, although proposals for doing so exist [23]. Here, a generalized method for implementing cooperative hybridization [31] is presented for dynamic DNA nanotechnology. In particular, cooperative hybridization find natural application in implementing robust and modular analog and digital nucleic acid logic circuits.

The Cooperative Hybridization Mechanism The cooperative hybridization mechanism is shown in Fig. 8-1. Two reactant strands, $T1$ and $T2$, hybridize simultaneously to multi-stranded complex $D1$ only when both are present. There are two parallel pathways for this reaction. In one, $T1$ first binds to $D1$ form intermediate complex I, which then reacts with $T2$ to form W and release product strand $P1$. In the other pathway, $T2$ binds first to $D1$ to form intermediate complex J. Individually, the hybridization of $T1$ or $T2$ to $D1$ is reversible and thermodynamically unfavorable; the release of product $P1$ upon the simultaneous hybridization of $T1$ and $T2$ drives this reaction forward.

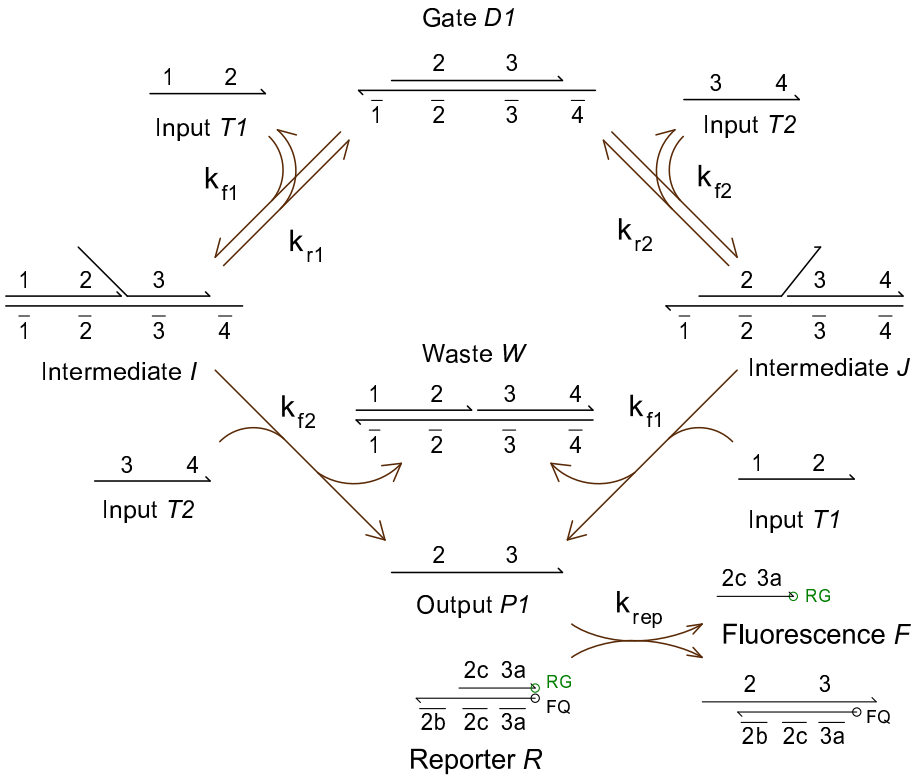


FIG. 8-1: The cooperative hybridization mechanism. Individual binding of $T1$ or $T2$ to gate $D1$ is thermodynamically unfavorable and kinetically reversible, but the simultaneous hybridization of both is thermodynamically favorable, and the release of product $P1$ inhibits the reverse reaction (dissociation of $T1$ and $T2$). The rate constant of $T1$ binding to $D1$ is assumed to be the same as that of $T1$ binding to intermediate J (k_{f1}), and the rate constant of $T2$ binding to $D1$ is assumed to be the same as that of $T2$ binding to I (k_{f2}). The reaction is assayed using fluorescent reporter complex R that reacts stoichiometrically with $P1$ to release unquenched fluorophore-labeled strand F. The bottom strand of $D1$ is named $U1$, and the bottom strand of R is named RU. The waste complex comprising $P1$ and RU is named Fw.

The cooperative hybridization mechanism can be expressed as the following reactions:

$$T1 + D1 \rightleftharpoons I$$

$$T2 + D1 \rightleftharpoons J$$

$$T2 + I \rightarrow P1 + W$$

$$T1 + J \rightarrow P1 + W$$

The net reaction of the system is thus:

$$T1 + T2 + D1 \rightarrow P1 + W$$

At operational conditions, the net reaction is strongly thermodynamically favorable.

There are two prominent features of this reaction. First, the equilibrium concentration of $P1$ will be roughly the minimum of the initial concentrations of $T1$, $T2$, and $D1$. As $P1$ is a strand with a different sequence than both $T1$ and $T2$, it can potentially participate in downstream reactions that $T1$ and $T2$ cannot. Second, because the individual reaction of $T1$ or $T2$ with $D1$ is thermodynamically unfavorable, very little of $T1$ or $T2$ is sequestered in I or J if only one of $T1$ or $T2$ is present. The equilbrium concentration of $T1$ or $T2$ in this case would be very near the total concentration of $T1$ or $T2$ in solution. These properties of the cooperative hybridization process allow a variety of useful dynamic nucleic acid devices and circuits.

Materials and Methods

DNA sequences and design. The oligonucleotide sequences used here were designed using a domain-based sequence design software [24] to possess minimal secondary structure and crosstalk (binding between unrelated domains). NUPACK [2] (http://www.nupack.org/) calculates there to be no more than 4 paired bases between any pair of strands at 25 °C, even at 1 μM concentration. Furthermore, the minimum free energy states of every individual strand was completely unstructures ($\Delta G^\circ = 0$). Thus, the domains we use can be approximated as structure-free. Substantial secondary structure is known to slow down branch migration and interfere with hybridization [25].

Standard free energy calculation. The standard free energies of complexes are needed in order to calculate the standard free energy of reactions, which in turn can be used to generate equilibrium and rate constants. NUPACK [2] was used to calculate the standard free energies of DNA complexes. NUPACK uses a number of different parameters in its calculations; the values used are detailed and justified below.

Domain	Sequence	Length (nt)
1	5'- CATCACTA -3'	8
2 = 2a:2b:2c	5'- CTATCATCACACATCTAT -3'	18
2a	5'- CTATCAT -3'	7
2b	5'- CACACAT -3'	7
2c	5'- CTAT -3'	4
3 = 3a:3b	5'- ACAACCACTTACTTCTTC -3'	18
3a	5'- ACAACCACTTACTT - 3'	14
3b	3'- CTTC -3'	4
4	5'- ATCTATCC -3'	8
5	5'- CTATCAT -3'	7
6	5'- CACACAT -3'	7
7	5'- CTATACAACCACTTACTT -3'	18
8	5'- CTTC -3'	4
9	5'- GCCATCAGAACTTAACCT -3'	18
10	5'- AACTC -3'	5
Strand	**Domain Composition**	**Length (nt)**
$T1$	1 2	26
$T2$	3 4	26
$P1$	2 3	36
$U1$	$\bar{4}$ $\bar{3}$ $\bar{2}$ $\bar{1}$	52
F	2c 3a	18
RU	$\bar{3}$a $\bar{2}$c $2\bar{\text{b}}$	25
$T3$	5 6 7 8	36
$S1$	9 10 ROX	23
$W2$	7 9	36
$U2$	RQ $\bar{1}\bar{0}$ $\bar{9}$ $\bar{7}\bar{6}$	48

TABLE 8-1: Domain and strand sequences. Barred domains are complements of the unbarred domains (e.g. $\bar{2}$ is complementary to 2), and their sequences are determined by those of their complements. For example, the 10 domain has sequence 5'- AACTC -3', so $1\bar{0}$ has sequence 5'- GAGTT -3'.

Temperature was set to 25 °C, as that was the temperature at which experiments were performed.

Salt concentration was set to 0.05 M Na^+ and 0.0115 M Mg^{2+}. In actuality, the experimental concentration of Na^+ is 0.002 M, but 0.05 M Na^+ was the lowest NUPACK allowed. However, since Mg^{2+} acts as the main counterion, it is likely that this difference does not significantly change the standard free energies.

The "dangles" parameter was set to ALL, so that dangles energies are incorporated for all bases flanking duplexes, regardless of whether it is paired. This is necessary because by default NUPACK does not incorporate the thermodynamics of coaxial stacks (such as at the nick in I, J, and W). The reason NUPACK does not by default include coaxial stacking

Complex	Strand Composition	Calculated Δ G°
$D1$	$P1$, $U1$	-49.27 + 2.38 = 46.9 kcal / mol
I	$P1$, $U1$, $T1$	-62.15 + 2·2.38 + 1.75= -55.6 kcal / mol
J	$P1$, $U1$, $T2$	-61.51 + 2·2.38 + 1.75 = -55.0 kcal / mol
W	$U1$, $T1$, $T2$	-70.69 + 2·2.38 = -65.9 kcal / mol
M	$U1$, $T1$	-34.89 + 2.38 = -32.5 kcal / mol
N	$U1$, $T2$	-35.81 + 2.38 = -33.4 kcal / mol

TABLE 8-2: Standard free energies of complexes shown in Fig. 8-1, 8-2, and 8-3. The standard free energies of these complexes were calculated using NUPACK [2], using the parameters 25 °C, 0.05 M Na^+, 0.0115 M Mg^{2+}, dangles = ALL. The standard free energies of all individual strands were 0 kcal / mol (completely unstructured). Because NUPACK yields Δ G° values that were calculated for mole fraction rather than molar, a corrective RT ln([H_2O]) = +2.38 kcal / mol term must be added for every strand in excess of 1. The intermediate states I and J each correspond to 19 isoenergetic branch migration states; the RT ln(19) = +1.75 kcal / mol term corrects for this state multiplicity.

thermodynamics is because they are still not well-understood: Pyshnyi et al. [26] report that the energetics of coaxial stacking near a nick depends significantly on the nearest neighbors (bases one away from the nick). Setting dangles = ALL allows partial compensation of the energetics at the nicks.

NUPACK reports ΔG° values that were calculated for mole fraction rather than molar; this means that equilibrium constants calculated using $\Delta G^\circ = -RT\ln(K)$ will be expressed in mole fraction rather than molar. To convert the NUPACK reported ΔG°, we add a corrective RT ln([H_2O]) = +2.38 kcal / mol term for every strand in excess of 1.

Intermediates I and J represent 3-stranded complexes on which branch migration is possible (domain 2 for I and domain 3 for J) Consequently, I and J each correspond to 19 isoenergetic branch migration states, and in the calculation of I and J's standard free energies, an extra RT ln(19)=1.75 kcal / mol term needs to be added in order to yield the proper equilibrium.

Table 8-2 shows the calculated free energies of the complexes.

Annealing. All annealing processes were performed with an Eppendorf Mastercycler Gradient thermocycler. The samples were brought down from 95 °C to 20 °C at a constant rate over the course of 90 minutes.

Complex purification. DNA oligonucleotides used in this study were purchased from Integrated DNA Technologies (IDT). Where applicable, fluorophores were attached by IDT

as well. Concentrations were determined from the measured absorbance at 260 nM using an Eppendorf Biophotometer and the calculated extinction coefficients [27].

Complexes $D1$ in Fig. 8-2 were further purified by non-denaturing (ND) polyacrylamide gel electrophoresis (PAGE) as follows: The $P1$ and $U1$ strands were prepared with nominally correct stochiometry at 20 μM and annealed. The samples were then run on 12% ND PAGE at 180 V for 6 hours.

The acrylamide (19:1 acrylamide:bis) was diluted from 40% acrylamide stock (Ambion). ND loading dye containing xylene cyanol FF in 50% glycerol was added to all samples, achieving final gycerol concentration of 10% by volume. Gels were run at 25 °C using a Novex chamber with external temperature bath.

The proper $D1$ band was cut out and eluted in 2 mL TE/Mg^{2+} buffer for 2 days. Purified complexes were quantitated by 260 nm absorbance measurement and calculated extinction coefficients. Yield was approximately 50%.

Buffer conditions. DNA oligonucleotides were stored in TE buffer (10 mM Tris · HCl pH balanced to 8.0, with 1 mM EDTA·Na_2, purchased as 100x stock from Sigma-Aldrich) at 4 °C. Directly preceding experiments, TE buffer with 62.5 mM $MgCl_2$ was added at 1:4 ratio to the sample, achieving a final $MgCl_2$ concentration of 12.5 mM (of which 1 mM is bound to EDTA). This buffer is henceforth known as "TE/Mg^{2+}" buffer. All experiments and purifications were performed at 25±0.5 °C, with temperature controlled using an external temperature bath.

Spectrofluorimetry studies. Spectrofluorimetry studies were done using a SPEX Fluorolog-3 (Horiba) with 1.6 mL 119-004F synthetic quartz cells (Hellma). The excitation were at 588 nm, while emissions were at 602 nm (optimal signal for ROX fluorophore). In all spectrofluorimetry experiments, the total reaction volume was 1.5 mL. For the rate constant characterization reactions (Fig. 8-2ABC), 4 nm band pass slits for used for both excitation and emission monochrometers; for all other experiments, 4 nm slits were used. In all experiments, data points were collected with an integration time of 10 seconds for every 60 second time-point.

Prior to each experiment, all cuvettes were cleaned thoroughly: each cuvette was washed

15 times in distilled water, once in 70% ethanol, another 5 times in distilled water, and finally once more in 70% ethanol.

For the slit size, concentrations, and times chosen, no measurable photobleaching was observed. All experimental results were within the linear regime of the spectrofluorimeter detector, according to specification sheets provided by the manufacturer.

Fluorescence normalization. Fluorescence is normalized so that 1 normalized unit (n.u.) of fluorescence corresponds to 1 nM of an unquenched fluorophore-labelled strand (F in Fig. 8-2 and 8-3, $S1$ in Fig. 8-4, 8-5, and 8-6). This normalization is based on the fluorescence levels of annealed samples: A negative control with $[R] = 30$ nM yielded the fluorescence of the quenched F strand (Rhodamine Green), and a positive control with $[R] = 30$ nM and $[P1] = 10$ nM yielded the fluorescence of the unquenched F strand. Similarly, a negative control with $[S1] = 5$ nM and $[U2] = 10$ nM yielded the fluorescence of the quenched $S1$ strand (ROX), and a positive control with $[S1] = 5$ nM yielded the fluorescence of the unquenched $S1$ strand.

Day-to-day and sample-to-sample variations in fluoroscence are estimated to be less than 5%.

Ordinary Differential Equation (ODE) simulations. Reaction simulations were run in Matlab using the "stiff" ode23s solver because of the difference in time scales of bimolecular and unimolecular reactions. The relative tolerance of the solver was set to 10^{-4}, and the absolute tolerance of the solver was set to 10^{-30} M. Sample code for simulating the system shown in Fig. 8-1 is shown in Text S1.

Parameter Fitting. Rate constants were fitted to experimental data using the "fminunc" function in Matlab to minimize the error between experimental data and the reaction model. The error between the data and the simulation is calculated as:

$$\text{Error} = \sum_{t,\text{traces}} (F_d(t) - F_m(t))^2$$

where $F_d(t)$ is the fluorescence value of the data at time t, and $F_m(t)$ is the fluorescence value predicted by the ODE model at time t. Sample code for fitting the k_{rep} rate constant

to the data in Fig. 8-2A is shown in Text S1.

The confidence intervals on the values of the rate constants were generated as the values of the rate constant at which the error score from the Reporter function is double that of the minimum. All other rate constants were kept constant at their best-fit values when generating the confidence interval on each rate constant.

Results

Characterizing cooperative hybridization. We begin by characterizing the rate constants for the experimental system. The k_{rep} rate constant for the $P1 + R \rightarrow F + Fw$ reporter reaction is measured in Fig. 8-2A to be $1.3 \pm 0.5 \cdot 10^7$ M^{-1} s^{-1}. This rate constant is significantly higher than those of similar reporter complexes [3, 18], though the reason for this is not understood. The rate constants of strand displacement reactions involving fluorophore and quencher-labelled DNA oligonucleotides generally cannot be predicted as reliably as those of unfunctionalized DNA [3, 28].

Th k_{f1} and k_{f2} rate constants are measured in Fig. 8-2B and 8-2C, respectively, to be $2.1 \pm 0.7 \cdot 10^6$ M^{-1} s^{-1} and $1.4 \pm 0.2 \cdot 10^6$ M^{-1} s^{-1}. These values are consistent with previous characterizations of strand displacement rate constants, which ranged from $3 \cdot 10^5$ M^{-1} s^{-1} to $6 \cdot 10^6$ M^{-1} s^{-1} for 8 nt toeholds, depending on the sequence of the toehold [3].

Fig. 8-2DEF show the experimental results on the full cooperative hybridization mechanism. This data was then used to fit the values of the k_{r1} and k_{r2} rate constants (0.54 ± 0.42 s^{-1} and 0.39 ± 0.27 s^{-1}, respectively). The reactions and rate constants simulated are shown in Table 8-3.

Table 8-3 also shows the calculated standard free energies and equilibrium constants of each intermediate reaction, using the standard free energies of DNA strands and complexes (Table 8-2). For comparison, the K_{eq} values can be calculated for the $T1 + D1 \rightleftharpoons I$ and $T2 + D1 \rightleftharpoons J$ reactions using the experimentally fitted rate constants. For the $T1 + D1 \rightleftharpoons I$ reaction, $K_{eq} = \frac{2.1 \cdot 10^6 \text{ M}^{-1} \text{ s}^{-1}}{0.54 \text{ s}^{-1}} = 3.9 \cdot 10^6$ M^{-1}. For the $T2 + D1 \rightleftharpoons J$ reaction, $K_{eq} = \frac{1.4 \cdot 10^6 \text{ M}^{-1} \text{ s}^{-1}}{0.39 \text{ s}^{-1}} = 3.6 \cdot 10^6$ M^{-1}. Thus, the equilibrium constant values calculated using experimentally fitted rate constants agreed to within an order of magnitude of those

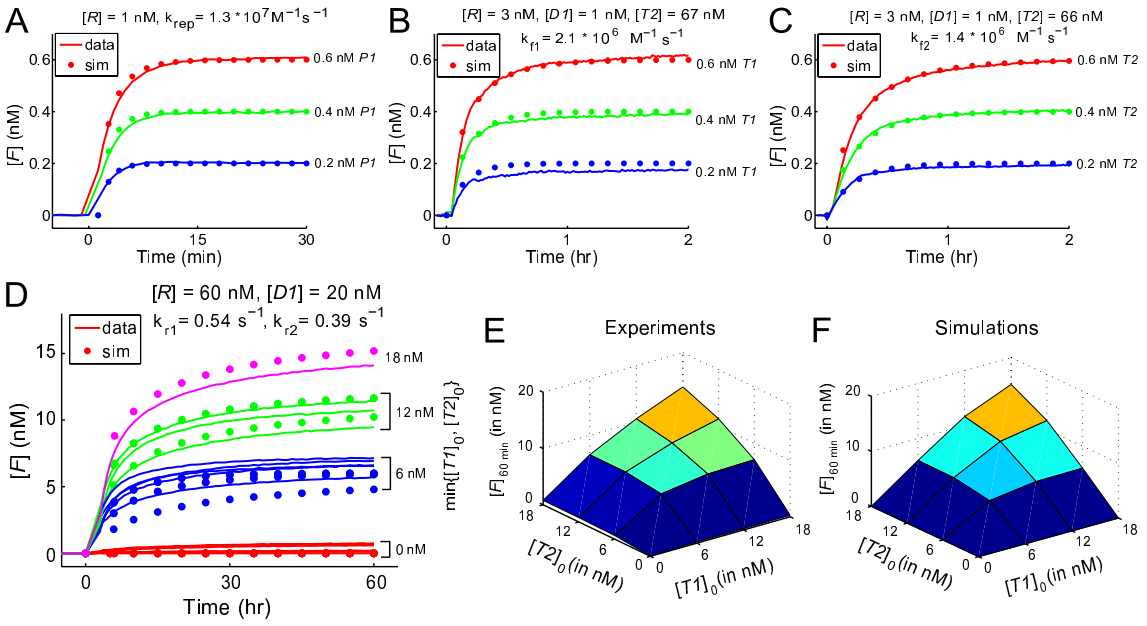


FIG. 8-2: Cooperative hybridization: modeling and experiments. **(A)** Characterization of the reporter rate constant k_{rep}. Reporter R was present in the cuvette initially, and $P1$ was added at time $t \approx 0$. Solid lines denote experimental data, and dotted lines denote ordinary differential equation (ODE) simulations of the reaction using the best fit rate constant $k_{rep} = 1.3 \cdot 10^7$ M^{-1} s^{-1} (shown in Figure). **(B)** Characterization of the $T1$ binding rate constant k_{f1}. Reporter R, gate $D1$ and input $T2$ were pre-mixed before the start of the reaction, and $T1$ was added at time $t \approx 0$. Due to the large excess of $T2$, it is expected that $D1$ is mostly present as intermediate J by time $t \approx 0$. ODE simulations are consistent with this assumption. The rate constant k_{f1} was fitted to the data shown here, using the value for k_{rep} fitted earlier. **(C)** Characterization of the $T2$ binding rate constant k_{f2}. Reporter R, gate $D1$ and input $T1$ were pre-mixed before the start of the reaction, and $T2$ was added at time $t \approx 0$. $D1$ is expected to be mostly presented as I. The rate constant k_{f2} was fitted to the data shown here, using the value for k_{rep} fitted earlier. **(D)** Characterization of the full cooperative hybridization mechanism. Reporter R and gate $D1$ were pre-mixed before the start of the reaction; $T1$ and $T2$ were added at time $t \approx 0$. Rate constants k_{r1} and k_{r2} were fitted to the data shown here, using the values for k_{rep}, k_{f1}, and k_{f2} fitted from earlier. **(E)** 3D summary of the experimental results in part (D). The final fluorescence is roughly linear in the minimum of $[T1]$ and $[T2]$; this allows cooperative hybridization to be an implementation of a logical AND gate, with $T1$ and $T2$ being the inputs, and $P1$ being the output. **(F)** 3D summary of simulation results, using fitted rate constants.

calculated based on the predicted thermodynamics of the DNA complexes (Table 8-3).

The agreement between ODE simulations and experimental data suggests that the cooperative hybridization mechanism functions as designed. The consumption of $T1$, the consumption of $T2$, and the production of $P1$ are all simultaneous, and stoichiometric with respect to each other. The equilibrium concentration of $P1$ becomes the minimum of $[T1]$ and $[T2]$. This means that cooperative hybridization could function as a logical AND gate, with $P1$ being output and $T1$ and $T2$ being inputs.

Thresholding. The $D1$ complex used in the previous section was purified by native PAGE

Reaction	Calc. ΔG° (25 °C)	Calc. K_{eq} (25 °C)
$T1 + D1 \overset{k_{f1}}{\underset{k_{r1}}{\rightleftharpoons}} I$	-8.7 kcal / mol	$2.4 \cdot 10^6$ M^{-1}
$T2 + D1 \overset{k_{f2}}{\underset{k_{r2}}{\rightleftharpoons}} J$	-8.1 kcal / mol	$8.6 \cdot 10^5$ M^{-1}
$T2 + I \overset{k_{f1}}{\rightarrow} P1 + W$	-10.3 kcal / mol	$3.5 \cdot 10^7$
$T1 + J \overset{k_{f2}}{\rightarrow} P1 + W$	-10.9 kcal / mol	$9.7 \cdot 10^7$
$P1 + R \overset{k_{rep}}{\rightarrow} F + Fw$	(unknown)	(unknown)
$k_{f1} = 2.1 \pm 0.7 \cdot 10^6$ M^{-1} s^{-1}	$k_{f2} = 1.4 \pm 0.2 \cdot 10^6$ M^{-1} s^{-1}	
$k_{r1} = 0.54 \pm 0.42$ s^{-1}	$k_{r2} = 0.39 \pm 0.27$ s^{-1}	
$k_{rep} = 1.3 \pm 0.5 \cdot 10^7$ M^{-1} s^{-1}		

TABLE 8-3: Reactions simulated in Fig. 8-2D. The middle and right columns shows the calculated Δ G$^\circ$ and K_{eq} of the reaction at 25 ∘C, based on the standard free energies shown in Table 8-2. The standard free energy of the $P1 + R \rightarrow F + Fw$ reaction cannot be accurately calculated: while fluorophore-quencher interactions are known to affect the thermdynamics of DNA hybridization [28], the free energy of these interaction has not been measured. The simulations use the best fit value of the rate constants; the errors bars on the rate constants are calculated as described in Materials and Methods: Parameter Fitting.

to ensure proper stoichiometry between $P1$ and $U1$. However, the $D1$ complex could also be prepared by mixing $P1$ with an excess of $U1$, without any subsequent purification; the excess $U1$ would act as a competitive threshold. $T1$ and $T2$ preferentially bind to $U1$ over $D1$, because $U1$ possesses more nucleation sites for initiating hybridization, and because the individual binding of $T1$ or $T2$ to $U1$ is thermodynamically favorable and kinetically irreversible (Fig. 8-3A).

Fig. 8-3B shows the experimental and simulation results of this system. The production of $P1$, and subsequently F, is sigmoidal in the minimum of $[T1]_0$ and $[T2]_0$. Tables 8-3 and 8-4 show the reactions involved in the thresholded system. The reactions shown in Table 8-4 are all hybridization reactions nucleated by any of over 20 base pairs, and are assumed to possess similarly high rate constants, k_h. The experimental results were used to fit $k_h = 3.1 \cdot 10^6$ M^{-1} s^{-1} and the concentration of $[U1] = 30$ nM. The concentration of $U1$ was measured by absorbance to be 25 nM. This discrepancy between the measured and fitted concentrations are likely due to a combination of synthesis impurities and inaccuracy in the extinction coefficients [29].

In the cooperative hybridization implementation of AND nucleic acid logic gates, thresholding provides a means of digital abstraction, preventing minor unintended side reactions from resulting in overall circuit error. However, this method of thresholding also yields an

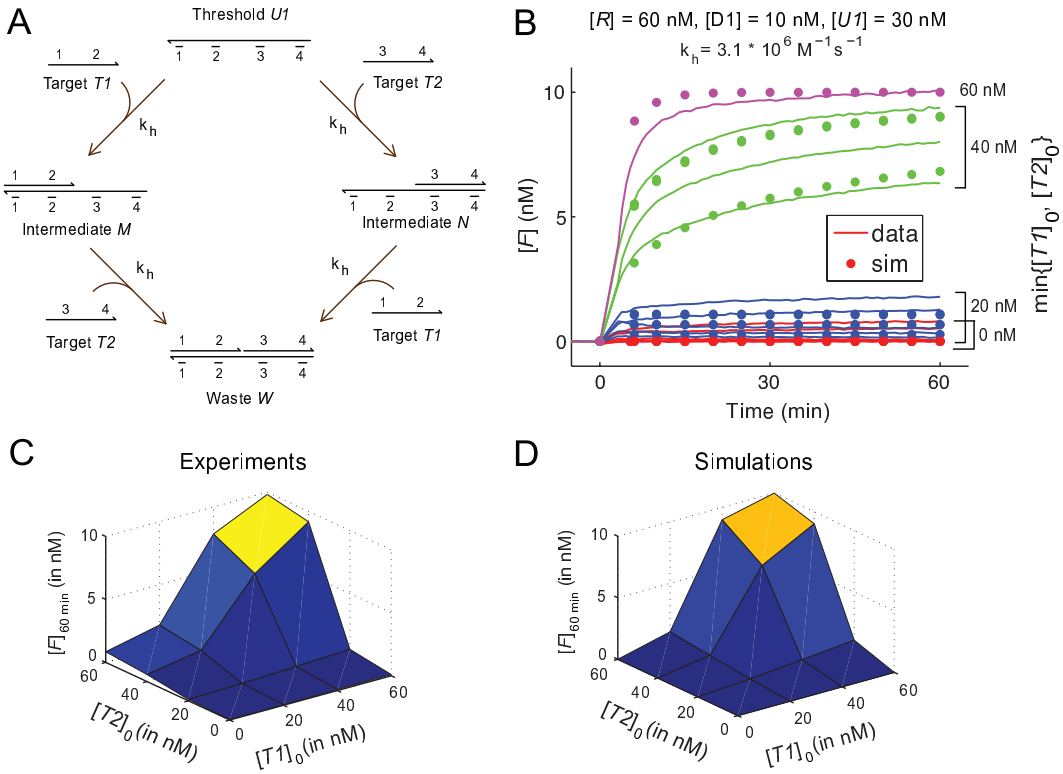


FIG. 8-3: Thresholding. **(A)** Excess bottom strand $U1$ serves as a threshold. $T1$ and $T2$ preferentially bind $U1$ over $D1$ because individual hybridization of $T1$ or $T2$ to $U1$ is thermodynamically favorable and kinetically irreversible. $T1$ and $T2$ are assumed to bind to $U1$, M, and N at the same rate constant k_h. **(B)** Experimental data and simulations on thresholded cooperative hybridization. R, $D1$, and $U1$ were pre-mixed before the start of the reaction; $T1$ and $T2$ were added at time $t \approx 0$. Rate constant k_h was fitted to the data shown here, using the value for k_{rep}, k_{f1}, k_{f2}, k_{r1}, and k_{r2} fitted from earlier. Additionally, the concentration of $U1$ was fitted here to be 30 nM; the calculated concentration of $U1$ (using absorbance and reported extinction coefficients) was 25 nM. Note that some simulation traces overlap (lower green dotted trace represent two almost identical simulation traces). **(C)** 3D summary of the experimental results in part (B). This system can serve as a thresholded AND logic gate. **(D)** 3D summary of simulation results, using fitted rate constants.

Reaction	Calc. Δ G°	Calc. K_{eq} (25 °C)
$T1 + U1 \stackrel{k_h}{\rightarrow} M$	-32.5 kcal / mol	$6.5 \cdot 10^{23}$ M^{-1}
$T2 + U1 \stackrel{k_h}{\rightarrow} N$	-33.4 kcal / mol	$3.0 \cdot 10^{24}$ M^{-1}
$T2 + M \stackrel{k_h}{\rightarrow} W$	-33.4 kcal / mol	$3.0 \cdot 10^{24}$ M^{-1}
$T1 + N \stackrel{k_h}{\rightarrow} W$	-32.5 kcal / mol	$6.5 \cdot 10^{23}$ M^{-1}
$k_h = 3.1 \pm 1.7 \cdot 10^6$ M^{-1} s^{-1}		

TABLE 8-4: Additional reactions simulated in Fig. 8-3 (all of the reactions shown in Table 8-3 were also simulated).

intrinsic signal loss: the concentration of output is necessarily lower than the concentrations of the inputs. For integrating many such thresholded AND logic gates into circuits, a method of signal amplification is also needed, such as through engineered DNA catalytic

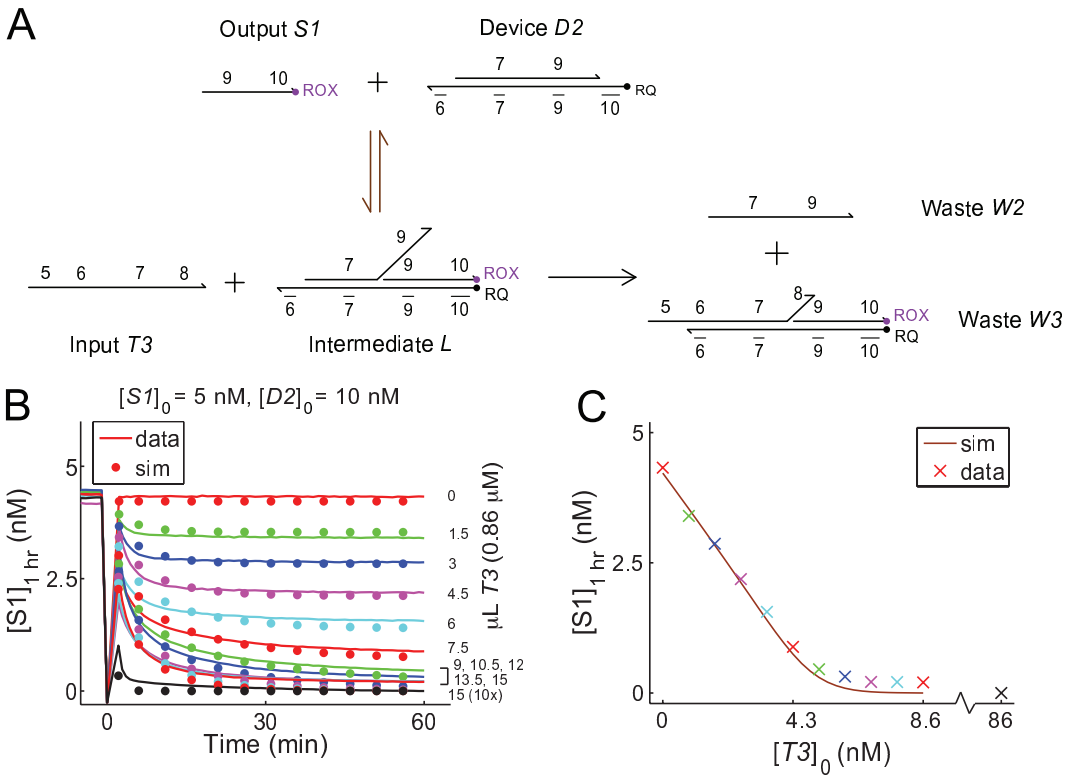


FIG. 8-4: Stoichiometric consumption via cooperative hybridization. **(A)** The cooperative hybridization mechanism ensures the stoichiometric consumption of its two oligonucleotide reactants. This mechanism for "annihilating" two oligonucleotides of independent sequence can be used as an analog subtraction circuit, or as an implementation of a logical NOT gate. In this implementation of a NOT gate, $T3$ acts as input and $S1$ serves as output; the concentration of free $S1$ decreases with the addition of $T3$. Only a subsequence of $T3$ (domains 7 and 8) partakes in the cooperative hybridization mechanism. The bottom strand of $D2$ is named $U2$. **(B)** Experimental data and simulations. $S1$ and $D2$ were present in solution initially, $T3$ was added at time $t \approx 0$. The sequences of $T3$ and $S1$ differ from those of $T1$ and $T2$, so the rate constants fitted in Fig. 8-2 do not apply here. The concentration of $T3$ was fitted here to be 0.86 nM; the calculated concentration of $T3$ (using absorbance and reported extinction coefficients) was 1.0 nM. See Table 8-4 for reactions simulations and rate constants fitted. **(C)** Summary of experimental and simulation results. The fluorescence after 1 hour of reaction is plotted against the concentration of $T3$. Note that in this subfigure (and only in this subfigure), the simulations are shown by the smooth line, while the experimental data are shown as colored crosses.

reactions [18, 31].

Stoichiometric consumption. To further verify the cooperative hybridization mechanism and to ensure that the consumption of $T1$ and $T2$ are stoichiometric and simultaneous, we characterize a slightly different cooperative hybridization system in this section (Fig. 8-4A). One of the two single-stranded reactants $S1$ is fluorophore-labeled, and the complementary complex $D2$ is quencher-labeled. This allows a direct assay of the instantaneous concentration of free $S1$.

Reactions	Rate constants
$T3 + D2 \underset{k_{r3}}{\overset{k_{f3}}{\rightleftharpoons}} K$	$k_{f3} = 4 \cdot 10^6$ M^{-1} s^{-1}
$S1 + D2 \underset{k_{r4}}{\overset{k_{f4}}{\rightleftharpoons}} L$	$k_{f4} = 4 \cdot 10^6$ M^{-1} s^{-1}
$S1 + K \overset{k_{f4}}{\rightarrow} W2 + W3$	$k_{r3} = 0.3$ s^{-1}
$T3 + L \overset{k_{f3}}{\rightarrow} W2 + W3$	$k_{r4} = 0.2$ s^{-1}

TABLE 8-5: Reactions simulated in Fig. 8-4. The data in Fig. 8-4 underconstrains the rate constants. Shown here are one set of rate constant values that generated reasonably good agreement between ODE simulations and experimental data. Similar qualities of fit can be attained by co-varying the 4 rate constants.

The other single-stranded reactant $T3$ possesses both a 5' and a 3' overhang that serve no function for the cooperative hybridization mechanism. The existence of these overhangs demonstrate that the cooperative hybridization mechanism can be used to target a subsequence of a longer nucleic acid, such as an mRNA.

Fig. 8-4B shows the results of this new system. The fluorescence of solution in the absence of $T3$ is between 4 and 4.5 nM, even though $[S1]_0 = 5$ nM. This is because a small fraction of the $S1$ is associated to $D2$ at equilibrium, and the corresponding intermediate L is low in fluorescence. The fluorescence (concentration of free $S1$) decreases as $[T3]_0$ increases, consistent with our understanding of the cooperative hybridization mechanism.

The length of domain 10 is only 5 nucleotides, rather than 7 or 8, because of the stabilizing effects of fluorophore-quencher binding on the thermodynamics of DNA hybridization [28]. However, the exact thermodynamics of the fluorophore-quencher binding at our experimental conditions have not been characterized, and it is not possible to predict the K_{eq} of the intermediate reactions as in the previous sections.

Table 8-5 shows the reactions simulated to generate the simulation traces shown in Fig. 8-4B and 8-4D. Five different parameters were fitted to the data in Fig. 8-4B (the 4 rate constants and the concentration of $T3$). Given the limited data on this system, the fit was underconstrained and it is likely that many different sets of rate constant values would have yielded fits of similar quality.

The stoichiometric consumption of the two single-standed reactants in cooperative hybridization ($T1$ and $T2$ in Fig. 8-1) can be interpreted as analog subtraction. If initially $[T1]_0 > [T2]_0$, then the equilibrium concentration of free $T1$ will be $[T1]_\infty \approx [T1]_0 - [T2]_0$

(assuming the equilibrum of $T1 + D1 \rightleftharpoons I$ favors the reactants). This property can be used to construct concentration comparison circuits, and may be useful for quantitation of nucleic acids [31].

Alternatively, the current system can be thought of as a logical NOT gate: $T3$ acts as the input species, and the equilibrium concentration of free $S1$ is low only when the initial concentration of $T3$ is high. Thresholding as described in the previous section can also be applied to this system, to allows an inverse-sigmoidal logical NOT response.

Cascaded reaction network exhibiting NAND logic. It has been shown that cooperative hybridization can be used to construct both logical AND and logical NOT gates. For practical applications involving non-trivial computations, it is necessary to show that these logic gates can be cascaded into circuits. In this section, the thresholded logical AND gate from Fig. 8-3 is cascaded with the logical NOT gate from Fig. 8-4 to form a cascaded circuit exhibiting logical NAND behavior.

To facilitate the cascading of these two gates, the sequences of $P1$ and $T3$ were designed to be identical. In a more general system, translator gates can be used to stoichiometrically convert oligonucleotides to other oligonucleotides of independent sequence [17, 30].

The cascaded NAND network behaves as expected: the free concentration of $S1$ decreases only when both $[T1]_0$ and $[T2]_0$ are higher than the threshold concentration $[U1]_0$ (Fig. 8-5B). ODE simulations using the rate constants previously fitted were able to qualitatively match experimental data traces without fitting any new parameters.

Thus far, all reactions have been tested in isolated solutions, in which no nucleic acids other than those needed were present. In order for the designed systems to function in either a biological setting or in a complex synthetic chemical network, however, all designed reactions must be robust to a background of nucleic acids that interact non-specifically.

Here, the robustness of the cooperative hybridizaton mechanism and the robustness of cascaded cooperative hybridization reaction networks are simultaneously tested by running the NAND reaction network in a solution of poly-N strands (Fig. 8-6). The poly-N strands are each 50 nt long, with a random (G, C, A, T) base at every position with roughly equal probability. As $4^{50} > 10^{23}$, it is likely that every strand in the poly-N mix is different in

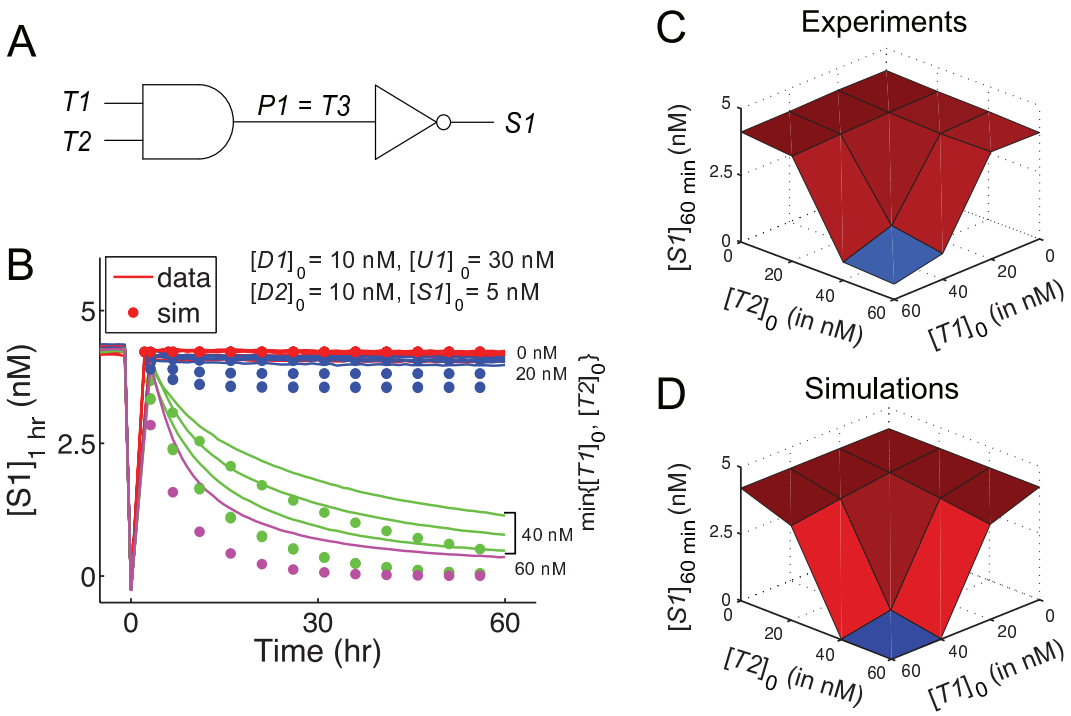


FIG. 8-5: Cascaded NAND logic network. **(A)** The $P1$ species from Fig. 8-2 and 8-3 was designed to be identical in sequence to $T3$ in Fig. 8-4. Consequently, the logical AND and the logical NOT gates can be directly cascaded to form a network exhibiting NAND logic. **(B)** Experimental data and simulations on the logical NAND cascaded network. $D1$, $U1$, $D2$, and $S1$ were present in solution initially, $T1$ and $T2$ were added at time $t \approx 0$. Importantly, no new parameters were fit to the data shown in this figure; rate constants simulated were fitted in previous figures. **(C)** 3D summary of the experimental results in part (B). **(D)** 3D summary of simulation results, using the previously fitted rate constants.

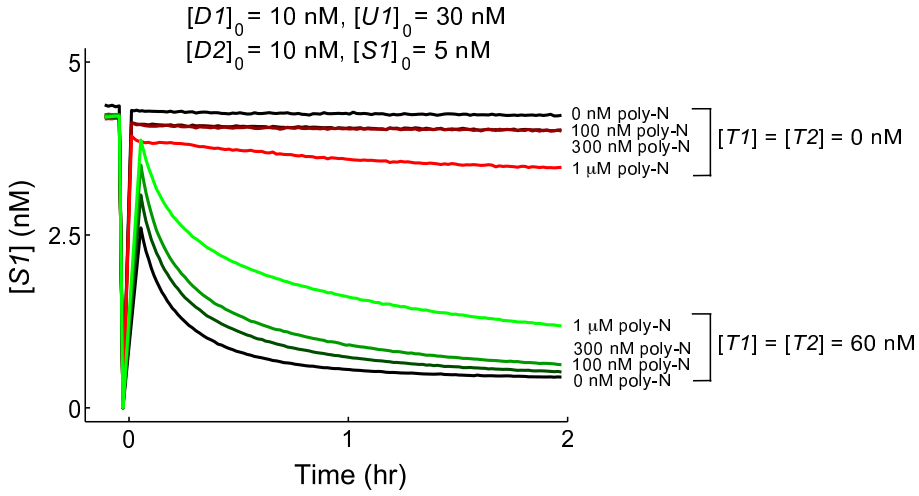


FIG. 8-6: Robustness to background molecules. Various amount of poly-N mix (50 nt oligonucleotides with a random distribution of G, C, A, and T bases at every position) was added the NAND reaction network. $D1$, $U1$, $D2$, and $S1$ were present in solution initially; $T1$, $T2$, and poly-N were added at time $t \approx 0$.

sequence.

As the amount of poly-N introduced increases, there is increased false positive; even in the absence of $T1$ and $T2$, the fluoroscence (corresponding to the concentration of free $S1$) is reduced over time. This, however, may be due to non-specific quenching of the fluorophore-labeled $S1$ strand: the guanine (G) nucleotide is known to quench many fluorophores, and the decreased observed fluorescence may be due to fleeting binding between single-stranded $S1$ and the poly-N mix strands.

In the case of excess $T1$ and $T2$, increasing amount of poly-N appear to slow the kinetics of the $[S1]$ decline, but the equilibrium fluorescence does not appear to be significantly changed. This agrees with our understanding of nucleic acid thermodynamics and kinetics: the poly-N mix could fleetingly bind to the single-stranded species $T1$, $T2$, $P1$, and $S1$, thus slowing the kinetics of the designed reactions, but it is unlikely that a significant fraction of the poly-N mix possesses enough complementarity to $S1$ to prevent $S1$ from binding to the perfectly complementary $U2$ strand.

Discussion

The thermodynamics and kinetics of the cooperative hybridization mechanism were examined in detail in this work, using a combination of experiments and simulations. Cooperative hybridization was seen implement a variety of circuit functions, both digital and analog, in addition to previously demonstrated applications in nucleic acid quantitation [31]. These can, in turn, be cascaded to form complex reaction networks that could potentially process and regulate biological nucleic acids.

Modeling. That the ODE simulations capture the behavior of the DNA molecules so well attests to the success of DNA thermodynamics and kinetics prediction parameters and algorithms. Nevertheless, several molecular phenomena need to be quantitatively characterized in order to predict the reaction kinetics of novel reaction mechanism from only strand sequence and reaction pathway information.

First, it is necessary to characterize the thermodynamics of various fluorophore-quencher interactions. While previous works [28] has characterized the effects of fluorophore-quencher interactions on the melting points of DNA duplexes, the standard enthalpy (ΔH°) and

entropy (ΔS°) of these interactions are not known. Consequently, at the present time it is not possible to predict the standard free energy (ΔG°) contribution of these interactions at various temperatures.

Second, it is necessary to better characterize the thermodynamics of coaxial stack interactions, the base-stacking energy at a nick in duplex DNA. The energetics of coaxial stacks have been shown to depend on the next nearest neighbors (bases 1 away from the nick) [26], so this is a more difficult problem than standard nearest neighbor base stacking parameters. Several different works characterizing the thermodynamics of coaxial stacks using only nearest neighbor information report values that differ from each other by up to 1.2 kcal/mol [26, 32, 33], which corresponds to the equilibrium constants differing by a factor of up to 7. Accurate coaxial stack thermodynamics are thus necessary for accurate quantitative prediction of rate constants from sequences.

Finally, the kinetics of DNA strand displacement reactions, for long toeholds, is bottlenecked by the kinetics of DNA hybridization, which in turn is bottlenecked by the nucleation of hybridization [34, 35]. The kinetics of nucleation DNA hybridization currently cannot be predicted from sequence, and varies by up to a factor of 20 depending on sequence [3]. Establishment of reliable elementary step-resolution simulations of nucleic acid reactions may yield an empirical method for determining the rate constant of hybridization [36].

Circuit applications. Cooperative hybridization was here demonstrated to implement analog subtraction and minimum functions. Both can serve as important roles in the analysis and evaluation of biological nucleic acids, in which over- and under-expression of an RNA relative to a standard expression level can be indicative of disease. Integrating these elements with amplification and readout could potentially allow the construction of nucleic acid devices that perform embedded computation within cells, such as determining cell state from microRNA concentrations.

These analog primitives can also serve useful functions in the construction of synthetic circuits for controlling nanoscale self assembly. In biology, cellular differentiation and development are often guided by the local relative expressions of different genes. Similarly, reaction-diffusion systems combining nucleic acid concentration gradients with concentra-

tion comparison circuits could lead to the formation of complex patterns and structures.

By combining these analog primitives with thresholding mechanisms, digital logic gates and circuits were constructed. Compared to previous implementations of nucleic acid logic gates and circuits based on strand displacement [17, 19], the major advantage of logical gates based on cooperative hybridization is the ease of sample preparation. As demonstrated in Fig. 8-3, there is no need for purification even to ensure stoichiometry, and Fig. 8-6 shows that these systems are robust to nucleic acid impurities in solution. This technical advantage reduces the labor needed to setting up these circuits, thereby facilitating the construction of more complex reaction networks.

However, like previous implementations, the cooperative hybridization implementation of AND and NOT gates possess certain limitations. Most notably, cooperative hybridization is a kinetically irreversible process, and logical values, once set, can never be reset. As a concrete example, in the NOT implementation (Fig. 8-4), the $S1$ output strand binds irreversibly to $D2$ and $T3$, and so the output value of the NOT gate can only be changed from ON to OFF, and never from OFF to ON. This practically means that the NOT gate is not dynamic–it cannot, for example, be used as part of a feedback circuit that requires multiple value changes, such as a three-ring oscillator composed of 3 serially cascaded NOT gates in a feedback loop.

Cooperative hybridization was here demonstrated to be a versatile and modular primitive. Systems based on cooperative hybridization were shown here to be cascaded to each other, and previous work has similarly shown that the cooperative hybridization reaction can be cascaded with strand displacement-based DNA catalysis reactions. Further integration with other dynamic DNA systems could allow useful high complexity nucleic acid devices that perform embedded computation, such as linear classifier circuits for distinguishing cell types.

Acknowledgements. The author is particular grateful to Erik Winfree for financial support for experimental reagents. This work was supported in part by NSF grants 0832824 and 0728703. DYZ is supported by the Fannie and John Hertz Foundation.

S1. Matlab code for simulating reactions and fitting rate constants

```
function dy = coop_hybridize(t, y)
%S + T1 -> I
%S + T2 -> J
%I -> S + T1
%J -> S + T2
%T1 + J -> P1
%T2 + I -> P1
%P1 + R -> F
%y = [S, T1, T2, I, J, P1, R, F]
krep = 1.3e7; %fitted earlier
kbind1 = 2.1e6; %fitted earlier
kbind2 = 1.4e6; %fitted earlier
krev1 = 0.54;
krev2 = 0.39;
dy = zeros(8,1);
dy(1) = -kbind1 * y(1) * y(2) - kbind2 * y(1) * y(3) + krev1 * y(4) + krev2 * y(5);
dy(2) = -kbind1 * y(1) * y(2) + krev1 * y(4) - kbind1 * y(2) * y(5);
dy(3) = -kbind2 * y(1) * y(3) + krev2 * y(5) - kbind2 * y(3) * y(4);
dy(4) = kbind1 * y(1) * y(2) - krev1 * y(4) - kbind2 * y(3) * y(4);
dy(5) = kbind2 * y(1) * y(3) - krev2 * y(5) - kbind1 * y(2) * y(5);
dy(6) = kbind1 * y(2) * y(5) + kbind2 * y(3) * y(4) - krep * y(6) * y(7);
dy(7) = -krep * y(6) * y(7);
dy(8) = krep * y(6) * y(7);
```

Note that the W and Fw species do not participate in any reactions, so the simulator does not explicitly calculate the concentrations of these complexes.

```
k0 = log(1E6);
scale0 = log(4e13);
[k, fval] = fminunc(@Reporter, [k0, scale0]);
```

The variable $k0$ shows an initial guess of the rate constant (set to 10^6 M^{-1} s^{-1} here), while $scale0$ shows an initial guess of the scaling constant (set to $4 \cdot 10^{13}$ counts per molar of unquenched fluorophore). To facilitate timely convergence on best-fit values, natural

logarithms of the rate constant values were used as inputs to fminunc. The fminunc function attempts to simultaneously fit both variables in order to minimize the output of the "Reporter" error function, shown below.

```
function err_func = Reporter(input)
data = load('/Users/daveyzhang/Desktop/work/expt/Fluorescence/20090119/001m.txt');
k = exp(input(1));
scalingconst = exp(input(2));
err_func = 0;
options = odeset('RelTol', 1e-4, 'AbsTol', 1e-30);
datasize = size(data, 1);
%0.2 nM data
t = data(7:datasize,1)-300;
y0 = [2e-10, 1e-9, 0];
[t, y2] = ode23s(@rdy, t, [k, y0], options);
ye = y2(:,4) * scalingconst;
for i = 7:size(data, 1)
    err_func = err_func + (ye(i-6) - (data(i,3)-data(5,3)+data(5,2)-data(i,2)))^2;
end

%0.4 nM data
t = data(7:datasize,1)-330;
y0 = [4e-10, 1e-9, 0];
[t, y4] = ode23s(@rdy, t, [k, y0], options);
ye = y4(:,4) * scalingconst;
for i = 7:size(data, 1)
    err_func = err_func + (ye(i-6) - (data(i,4)-data(5,4)+data(5,2)-data(i,2)))^2;
end

%0.6 nM data
t = data(7:datasize,1)-360;
y0 = [6e-10, 1e-9, 0];
[t, y6] = ode23s(@rdy, t, [k, y0], options);
ye = y6(:,4) * scalingconst;
for i = 7:size(data, 1)
    err_func = err_func + (ye(i-6) - (data(i,5)-data(5,5)+data(5,2)-data(i,2)))^2;
end
```

[1] SantaLucia, J. & Hicks, D. The Thermodynamics of DNA Structural Motifs. *Ann. Rev. Biochem.* **33,** 415-440 (2004).

[2] Dirks, R. M., Bois, J. S., Schaeffer, J. M., Winfree, E. & Pierce, N. A. Thermodynamic Analysis of Interacting Nucleic Acid Strands. *SIAM Review* **49,** 65-88 (2007).

[3] Zhang, D. Y. & Winfree, E. Control of DNA Strand Displacement Kinetics Using Toehold Exchange. *J. Am. Chem. Soc.* **131,** 17303-17314 (2009).

[4] Yurke, B. & Mills, A. P. Using DNA to Power Nanostructures, *Genetic Programming and Evolvable Machines* **4,** 111-122 (2003).

[5] Winfree, E., Liu, F., Wenzler, L. A. & Seeman, N. C. Design and self-assembly of two-dimensional DNA crystals. *Nature* **394,** 539-544 (1998).

[6] Dirks, R. M. & Pierce, N. A. Triggered Amplification by Hybridization Chain Reaction. *Proc. Nat. Acad. Sci.* **101,** 15275-1278 (2004).

[7] Rothemund, P. Folding DNA to create nanoscale shapes and patterns. *Nature* **440,** 297-302 (2006).

[8] Yin, P., Choi, H. M. T., Calvert, C. R. & Pierce, N. A. Programming biomolecular self-assembly pathways. *Nature* **451,** 318-322 (2008).

[9] Rothemund, P. W. K., Papadakis, N. & Winfree, E. Algorithmic self-assembly of DNA Sierpinski triangles. *Plos Biol.* **2,** 2041-2053 (2004).

[10] Zheng, J., Birktoft, J. J., Chen, Y., Wang, T., Sha, R., Constantinou, P. E., Ginell, S. L., Mao, C. & Seeman, N. C. From molecular to macroscopic via the rational design of a self-assembled 3D DNA crystal. *Nature* **461,** 74-77 (2009).

[11] Douglas, S. M., Dietz, H., Liedl, T., Hogberg, B., Graf, F. & Shih, W. M. Self-assembly of DNA into nanoscale three-dimensional shapes. *Nature* **459,** 414-418 (2009).

[12] Andersen, E. S., Dong, M., Nielsen, M. M., Jahn, K., Subramani, R., Mamdouh, W., Golas, M. M., Sander, B., Stark, H., Oliveira, C. L. P., Pedersen, J. S., Birkedal, V., Besenbacher, F., Gothelf, K. V. & Kjems, J. Self-assembly of a nanoscale DNA box with a controllable lid. *Nature* **459,** 73-76 (2009).

[13] Yurke, B., Turberfield, A. J., Mills, A. P., Simmel, F. C. & Neumann, J. L. A DNA-fuelled molecular machine made of DNA. *Nature* **406,** 605-608 (2000).

[14] Omabegho, T., Sha, R. & Seeman, N. C. A Bipedal DNA Brownian Motor with Coordinated Legs. *Science* **324,** 67-71 (2009).

[15] Bath, J & Turberfield, A. J. DNA nanomachines. *Nature Nanotech.* **2,** 275-284 (2007).

[16] Green, S., Bath, J. & Turberfield, A. J. Coordinated Chemomechanical Cycles: A Mechanism for Autonomous Molecular Motion. *Phys. Rev. Lett. 101,* 238101 (2008).

[17] Seelig, G., Soloveichik, D., Zhang, D. Y. & Winfree, E. Enzyme-free nucleic acid logic circuits. *Science* **314,** 1585-1588 (2006).

[18] Zhang, D. Y., Turberfield, A. J., Yurke, B. & Winfree, E. Engineering entropy-driven reactions and networks catalyzed by DNA. *Science* **318,** 1121-1125 (2007).

[19] Frezza, B. M., Cockroft, S. L. & Ghadiri, M. R. Modular multi-level circuits from immobilized DNA-based logic gates. *J. Am. Chem. Soc.* **129,** 14875-14879 (2007).

[20] Seelig, G., Yurke, B. & Winfree, E. Catalyzed Relaxation of a Metastable DNA fuel. *J. Am. Chem. Soc.* **128,** 12211-12220 (2006).

[21] Green, S. J, Lubrich, D. & Turberfield, A. J. DNA hairpins: fuel for autonomous DNA devices. *Biophysical Journal* **91,** 2966-2975 (2006).

[22] Zhang, D. Y. & Winfree, E. Dynamic allosteric control of noncovalent DNA catalysis reactions. *J. Am. Chem. Soc.* **130,** 13921-13926 (2008).

[23] Soloveichik, D., Seelig, G. & Winfree, E. DNA as a universal substrate for chemical kinetics. *Proc. Nat. Acad. Sci.* **107,** 5393-5398 (2010).

[24] Zhang, D. Y. Towards Domain-based Sequence Design for DNA Strand Displacement Reactions. (submitted to the 16th International Conference on DNA Computing and Molecular Programming, 2010).

[25] Gao, Y., Wolf, L. K. & Georgiadis, R. M. Secondary structure effects on DNA hybridization kinetics: a solution versus surface comparison. *Nuc. Acids Res.* **34,** 3370-3377 (2006).

[26] Pyshnyi, D. V. & Ivanova, E. M. The influence of nearest neighbours on the efficiency of coaxial stacking at contiguous stacking hybridization of oligodeoxyribonucleotides. *Nucleosides, Nucleotides, and Nucleic Acids* **23**, 1057-1064 (2004).

[27] Puglisi, J. D. & Tinoco, I. Absorbance melting curves of RNA. *Methods in Enzymology* **180**, 304-325 (1989).

[28] Marras, S. A. E., Kramer, F. R. & Tyagi, S. Efficiencies of fluorescence resonance energy transfer and contact-mediated quenching in oligonucleotide probes. *Nuc. Acids Res.* **30,** e122 (2002).

[29] Owczarzy, R., Moreira, B. G., You, Y., Behlke, M. A., Walder, J. A. Predicting stability of DNA duplexes in solutions containing magnesium and monovalent cations. *Biochemistry* **47,** 5336-5353 (2008).

[30] Picuri, J. M., Frezza, B. M. & Ghadiri, M. R. Universal translators for nucleic acid diagnosis. *J. Am. Chem. Soc.* **131,** 9368-9377 (2009).

[31] Zhang, D. Y. Digital Sensors for Comparing Oligonucleotide Concentrations. (submitted, 2010).

[32] Peyret, N. *Prediction of Nucleic Acid Hybridization: Parameters and Algorithms.* Doctoral thesis, Wayne State University (2000).

[33] Pyshnyi, D. V. & Ivanova, E. M. Thermodynamic parameters of coaxial stacking on stacking hybridization of oligodeoxyribonucleotides. *Russian Chemical Bulletin* **51**, 1145-1155 (2002).

[34] Bloomfield, V. A.; Crothers, D. M.; Tinoco, I. Jr. *Nucleic Acids: Structures, Properties, and Functions* **2000,** University Science Books, Sausalito, CA.

[35] Jo, J. J., Kim, M. J., Son, J. T., Kim, J. & Shin, J. S. Single-fluorophore monitoring of DNA hybridization for investigating the effect of secondary structure on the nucleation step. *Biochemical and*

Biophysical Research Communications. **385,** 88-93 (2009).

[36] Flamm, C., Fontana, W., Hofacker, I. L & Schuster, P. RNA folding at elementary step resolution. *RNA* **6,** 325-338 (2000).

Chapter 9: Fixed Gain and Linear Classification

Cells are more than just their DNA–that's why brain cells and lung cells and cancerous lung cells are different, despite having more or less the same genome. For the most part, gene expression (the quantity of mRNA and protein that a gene eventually produces) determine the development and identity of cells: some genes are expressed in high quantities, others in low quantities, and others not at all. Different combinations of gene expressions lead to different cell types.

If we can design molecular reaction networks that can analyze gene expression at a molecular level to identify cell type and state, we can potentially construct nucleic acid devices that react differently in different cells. For example, these devices could activate programmed cell death pathways in cells that are identified as being cancerous or otherwise diseased. Professor Niles Pierce has constructed reaction networks that triggers apoptosis via the protein kinase R (PKR) pathway in the presence of an oligonucleotide of specific sequence.

It is rare, however, for a certain nucleic acid to be expressed only in cancer or other diseased cells and be completely absent in healthy cells. In order to design the next generation of programmable therapeutic biomolecules, it is necessary to be able to reliably distinguish cells based on the combination of over- and under-expressions of various nucleic acids.

In the work below, Georg and I propose and analyze a scheme for classifying solutions based on the concentrations of various nucleic acid molecules. In this work, we're restricting ourselves to linear classification–where linear combinations of the concentrations of various nucleic acids are sufficient to determine the type and status of the solution (or cell). Recent research by Lu et al. shows that a large variety of cancers can be classified linearly; it is possible that by analyzing enough nucleic acids in parallel, linear classification would be sufficient to distinguish cells for all practical purposes.

This work was submitted to the 16th annual DNA Computing and Molecular Programming Conference as:

Zhang, David Yu; Seelig, Georg. "DNA-based Fixed Gain Amplifiers and Linear Classifier Circuits."

DNA catalysts have been developed as methods of amplifying single-stranded nucleic acid signals. The maximum turnover (gain) of these systems, however, often varies based on strand and complex purities, and has so far not been well-controlled. Here we introduce methods for controlling the asymptotic turnover of strand displacement-based DNA catalysts and show how these could be used to construct linear classifier systems.

Introduction

DNA nanotechnology has utilized the specific binding properties [1] and the well-understood thermodynamics [2] and kinetics [5, 49] of nucleic acids to construct dynamic cascaded reactions, such as logic gates and circuits [10–12, 46], motors [6, 23], and amplification mechanisms [16, 17, 19–22, 25].

DNA devices can operate in complex biochemical environments and can be programmed to specifically interact with biological nucleic acids such as messenger RNA (mRNA) or microRNA (miRNA). DNA circuits could be used to develop novel point-of-care diagnostic devices that integrate detection with analysis and do not require complex laboratory

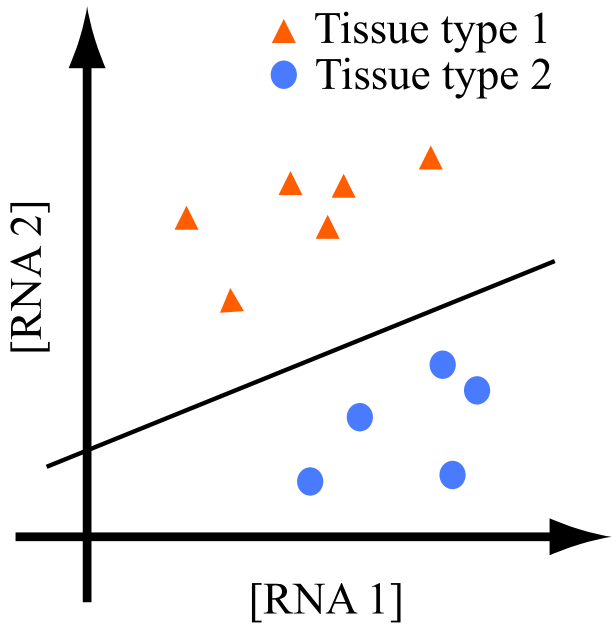


FIG. 9-1: Sketch of a hypothetical two-gene classifier. Samples from two different tissue can be clearly distinguished based on the expression profiles of two RNA molecules.

equipment. It has even been suggested to use DNA devices as "smart therapeutics" that operate inside living cells and integrate detection of specific disease markers with the activation of a therapeutic response based on the RNA interference pathway [7, 15], on antisense oligonucleotides [14] or ribozymes.

Such applications require nucleic acid circuitry that can reliably identify a specific disease state. Characteristic RNA markers that could serve as inputs to a DNA analytic circuit have been identified for many diseases. However, it is often not sufficient to simply detect the presence or absence of a set of RNA markers. Instead the classifiers that distinguish a disease tissue from healthy tissue (or other disease tissues) are often complex functions of the concentrations of multiple RNA markers (see Refs. [8, 9] for examples of microRNA-expression based classifiers of varying complexity).

Here we propose a molecular implementation for a specific class of classifiers, namely linear classifiers. The classifier circuit computes a linear combination with arbitrary (positive or negative) weights on a set of inputs (e.g. RNA molecules) and compares the result to a threshold value. Fig. 9-1 shows a highly simplified sketch of a linear two-gene classifier: the line separating the two different tissue types is given by an equation of the form

$\alpha_1[RNA_1] + \alpha_2[RNA_2] = K$. Given a sample of unknown origin, we can now classify it as tissue type 1 or 2 based on a measurement of two RNAs. Unlike in the more conventional case where the expression of each RNA is individually measured and the linear classification analysis is performed *in silico*, here both detection and analysis are done on the molecular level, allowing *in situ* and *in vivo* applications.

Previous DNA logic gates and circuits were mostly designed for a situation where inputs can be represented as Boolean variables and are either present at a high concentration or completely absent [10, 46]. This does not necessarily require the original inputs to be at a specific level; DNA-based signal restoration units consisting of a threshold gate and an amplifier can be used to restore an input with an arbitrary concentration to the expected logical TRUE or FALSE values. Still, the digital nature of such circuits is inherently incompatible with classification problems, in which the relative amounts of inputs determines the value of the final output. The fixed gain amplification methods presented here allow reliable tuning of analog signals encoded in the concentrations of nucleic acids.

Fixed gain amplifiers: lowering catalytic turnover

One key component of the proposed linear classifier is a DNA-based catalytic amplifier, that allows one signal-stranded nucleic acid to specifically produce or release many single-stranded nucleic acid molecules of independent sequence. Importantly, this amplifier needs to have a finite and controllable gain α such that each input on average releases α copies of the output. Such a finite gain amplifier would be useful not only in a linear classifier, where each detected RNA species is assigned a different weight, but could also be used for pre-amplification of a set of low-concentration inputs while maintaining their relative concentrations.

Existing DNA amplifiers have an intrinsically finite turnover; strand displacement-based nucleic acid catalysts typically convert on the order of 10-100 substrates before being inactivated [16, 20]. Inactivation is most likely due to defective substrate complexes or fuels [16, 19, 20]. The details of the inactivation process depend on the specifics of the amplifier design, but it seems likely that imperfectly synthesized DNA strands are a major culprit.

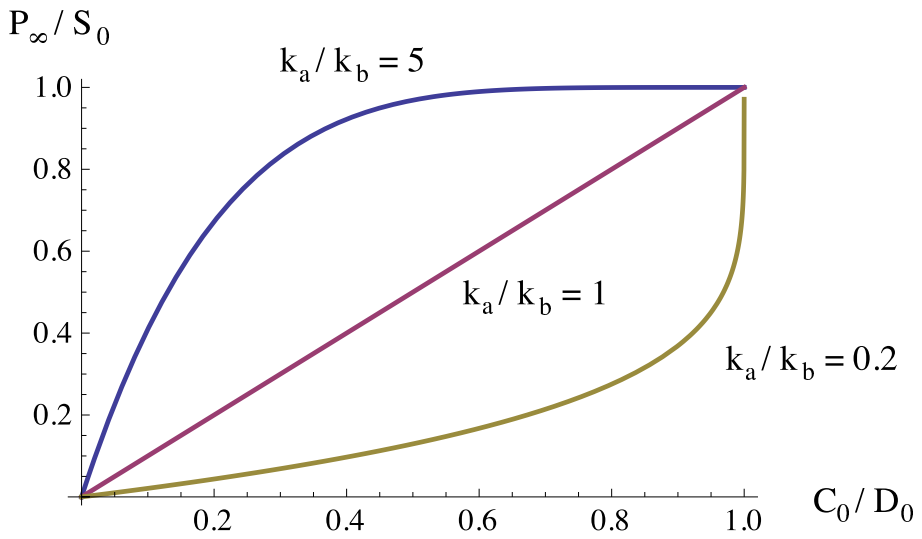


FIG. 9-2: Output produced in a catalytic reaction with competitive inhibition as a function of the catalyst concentration. Final product P_∞ is scaled by initial concentration of substrate S_0. Initial catalyst concentration C_0 is measured in units of inhibitor concentration I_0. We obtain different I/O characteristics depending on the ratio between the rate constants k_a and k_b for the catalytic and the competitive reaction.

In practice, the maximal turnover obtained seems to depend strongly on sequence, purification procedures, strand orientation and similar experimental and design details. Therefore, while the gain is finite, it can be characterized for any particular system.

The question then becomes if, starting from an arbitrary but high turnover, we can lower the turnover controllably to a fixed value. Given the intrinsic turnover of a catalytic system, intuitively it seems clear that we can lower the turnover further either by increasing the fraction of imperfect substrate or through addition of an alternative competitive inhibitor that irreversibly binds to the catalyst. However, it may be less intuitive how to best adjust the turnover to any specific desired value.

To address the question of how to control turnover we first consider a simple model for a catalytic reaction with competitive inhibition. Afterwards, we simulate a specific DNA implementation using measured reaction parameters. A catalytic reaction in the presence of an impurity can be modeled as:

$$C + S \xrightarrow{k_a} C + P \tag{1}$$

$$C + D \xrightarrow{k_b} \emptyset \tag{2}$$

In the first reaction a catalyst C transforms a substrate S into a product P. The rate constant for this reaction is k_a. The catalyst can also participate in a second, unproductive reaction with an inhibitor (or damper) D. This reaction proceeds at a rate constant k_b.

The differential equations resulting from this model can be integrated with initial conditions $C(0) = C_0$, $S(0) = S_0$, $D(0) = D_0$ and $P(0) = 0$. Solving for the product $P(t)$ we get

$$P(t) = S_0 - S_0 \left(\frac{1-\rho}{1-\rho e^{k_b \Delta t}} \right)^{k_a/k_b} \tag{3}$$

where we introduced the ratio $\rho = C_0/D_0$ and the difference $\Delta = C_0 - D_0$ of the initial amounts of catalyst and inhibitor.

In an ideal system without competitive inhibition the final product concentration is always equal to the initial concentration of substrate. Given enough time the catalyst will convert all substrate into product. In a system with competitive inhibition this is not necessarily true. The final amount of product produced in that case can be computed by taking the limit $t \to \infty$ in Eq. 3:

$$\lim_{t\to\infty} P(t) = P_\infty = \begin{cases} S_0, & C_0 \geq D_0 \\ S_0 - S_0 \left(1-\rho\right)^{k_a/k_b}, & C_0 < D_0 \end{cases} \tag{4}$$

Not surprisingly, if we start out with more catalyst than inhibitor, the reaction will eventually go to completion. The opposite limit is more interesting.

First, consider the case where the rate for the catalytic reaction is much faster than the inhibition reaction, $k_a > k_b$ (blue trace in Fig. 9-2). In this case the inhibitor has a relatively minor effect that is most pronounced at low concentrations of catalyst compared to the inhibitor.

In the limit where the catalytic reaction occurs at exactly the same rate as the inhibitory reaction, i.e. $k_a = k_b$ (red trace in Fig. 9-2) Eq. 4 predicts that the final amount of product is linear in the initial amount of catalyst, i.e. $P_\infty = \alpha C_0$ where $\alpha = S_0/D_0$. That is, by adjusting the relative concentration of substrate to inhibitor we can get any finite gain we

$S + C \underset{k_2}{\overset{k_1}{\rightleftharpoons}} I1 + SP$	$k_0 = 5\ \mathrm{M}^{-1}\ \mathrm{s}^{-1}$
$I1 + F \overset{k_2}{\rightarrow} I2 + OP$	$k_1 = 2.7 \cdot 10^5\ \mathrm{M}^{-1}\ \mathrm{s}^{-1}$
$I2 \underset{k_1}{\overset{k_3}{\rightleftharpoons}} W + C$	$k_2 = 1.1 \cdot 10^6\ \mathrm{M}^{-1}\ \mathrm{s}^{-1}$
$S + F \overset{k_0}{\rightarrow} OP + SP + W$	$k_3 = 1.1 \cdot 10^{-2}\ \mathrm{s}^{-1}$
$I1 + Fb \overset{k_2}{\rightarrow} X + OP$	$k_{rep} = 4 \cdot 10^5\ \mathrm{M}^{-1}\ \mathrm{s}^{-1}$
$S + Fb \overset{k_0}{\rightarrow} OP + SP + W2$	
$C + W2 \overset{k_1}{\rightarrow} X$	
$C + D \overset{k_1}{\rightarrow} X$	
$A + C \underset{k_2}{\overset{k_1}{\rightleftharpoons}} IA1 + SP$	
$IA1 + F \overset{k_2}{\rightarrow} I2 + C$	
$I2 \underset{k_1}{\overset{k_3}{\rightleftharpoons}} W + C$	
$A + F \overset{k_0}{\rightarrow} C + SP + W$	
$IA1 + Fb \overset{k_2}{\rightarrow} X + C$	
$A + Fb \overset{k_0}{\rightarrow} C + SP + W2$	

Table 9-1: Reactions simulated in Fig. 9-4.

need.

The situation where the rate for the inhibitor reaction is faster than the rate for the catalytic reaction is also interesting. In that case, the amount or product is sub-linear in the initial amount of catalyst for $C_0 < D_0$ but reaches a fixed value S_0 in the opposite regime. The concentration of the competitive inhibitor I therefore acts as a threshold for the catalytic reaction. Such a threshold element is useful for reliable signal propagation for example in the context of chemical digital circuits.

We now turn to a specific DNA implementation of such a system. Our implementation is based on the entropy-driven catalytic amplifier of Ref. [16] which was further characterized in Ref. [19]. Turnover for this amplifier was measured to be about 100. The reaction mechanism for this system including the side reactions leading to intrinsically finite turnover is shown in Fig. 9-3A. A reaction between catalyst strand and substrate relies on toehold mediated strand displacement. As a competitive inhibitor we here propose to use a damper DNA gate that irreversibly binds the catalytic input (Fig. 9-3B). In order to match the reaction rate constants of the catalyst with this inhibitor to that of the catalyst with the active substrate we simply choose the toeholds for both reactions to be identical.

In order to verify the predictions from our simple model Eq. 1 we simulated the full catalytic system of Ref. [16] with a parallel inhibitory reaction using the measured rate

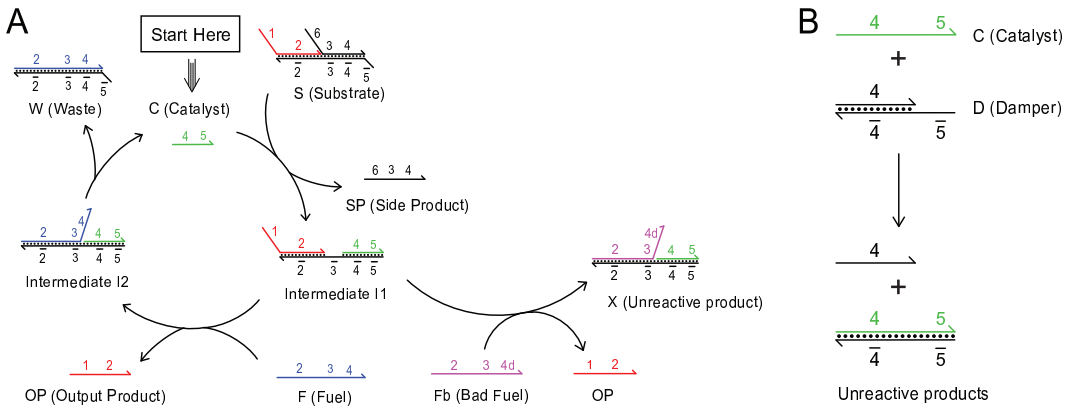


FIG. 9-3: Methods for tuning catalytic turnover. **(A)** DNA amplification via catalysis, adapted from Zhang et al. [19]. Catalyst strand C reacts with S to form side product SP and intermediate $I1$, the latter of which subsequently reacts with F to release output product OP, waste W, and catalyst C. However, a small fraction of bad fuel with deletions and/or degradation near the 3' end, denoted as Fb, will bind to intermediate $I1$ to form an unreactive product X, thus permanently trapping catalyst C and reducing the observed catalytic turnover of the reaction. The ratio $\frac{[Fb]}{[F]+[Fb]}$ was estimated to be 0.01 for HPLC-purified fuel strands [19]. **(B)** The catalytic turnover of the reaction can be tuned to be lower via the addition of the damping complexes D. Because C binds by the toehold to D as to S, it is assumed that this rate constant is identical in value to that of k_1. **(C)** Schematic of a generalized catalytic reaction with arbitrary control over turnover. In the original work on entropy-driven DNA catalysts [16], it was shown that the catalytic reaction can be made autocatalytic by using an alternative substrate A, which releases as product an molecule identical to the catalyst. Turnover can be increased above the limit set by fuel purity with autocatalytic substrate.

constants and reaction intermediates. The model is given in Table 9-1 and resulting data is shown in Fig 9-4A. As expected from our model, the final fluorescence depends linearly on the concentration of the damper gate.

Fixed gain amplifiers: increasing catalytic turnover

The turnover of a catalytic reaction can be increased above the intrinsic limit set by defective oligonucleotides. It seems clear that it should be possible to compensate for the loss of catalyst in an unproductive reaction through the production of an extra catalyst in a parallel autocatalytic reaction that proceeds at the same rate. A simple model motivated by this intuition is

$$C + S \xrightarrow{k_a} C + P, \tag{5}$$

$$C + D \xrightarrow{k_b} \emptyset, \tag{6}$$

$$C + A \xrightarrow{k_b} 2C. \tag{7}$$

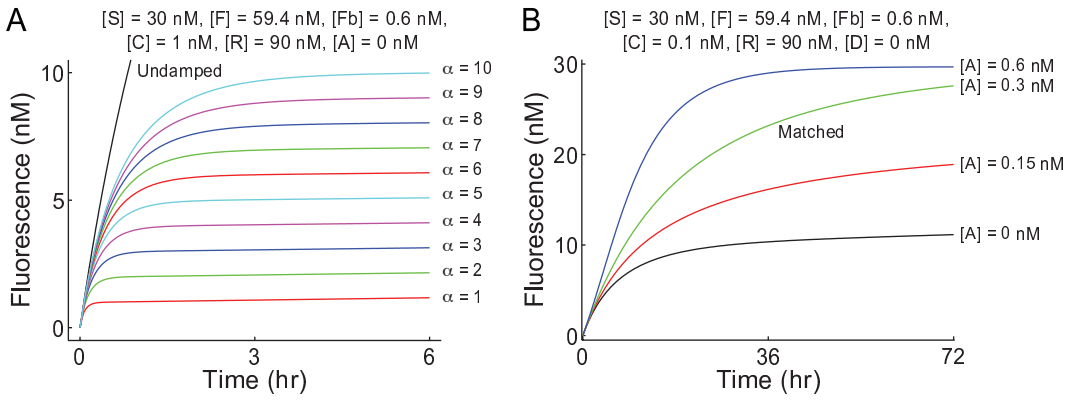


FIG. 9-4: Modulating turnover. **(A)** Simulations of the entropy-driven catalyst system with damper. [19]. Various amounts of D were present to achieve the fixed turnover η shown, with $[D] = \frac{30}{\alpha} - 0.3$ nM. See Table 9-1 for the full set of simulated reactions. **(B)** Simulations of the entropy-driven catalyst system with autocatalytic substrate. At $[A] = 0.3$ nM, the increase of catalyst due the autocatalytic substrate nearly matches the decrease of the catalyst due to bad fuel. With lower concentrations of A, asymptotic turnover is limited. With higher concentrations of A, the reaction adopts autocatalytic characteristics, and becomes less sensitive to the initial concentration of the catalyst.

Here A is the substrate for the autocatalytic reaction which is present initially at a concentration $A(0) = A_0$. With the same initial conditions as above we can solve the resulting differential equation. The final product as a function of time then is

$$P(t) = S_0 - S_0 \left(\frac{1-\sigma}{1-\sigma \mathrm{e}^{k_a \Gamma t}} \right)^{k_a/k_b}, \tag{8}$$

where $\Gamma = C_0 + A_0 - D_0$ and $\sigma = C_0/(D_0 - A_0)$. The result is therefore of exactly the same form as Eq. 3 if we make the substitution $D_0 \to D_0 - A_0$. In the special case where the initial concentrations of the inhibitor D and the substrate A for the autocatalytic reactions are the same, i.e. $A_0 = D_0$, these reactions cancel each other out and $P(t) = S_0 \exp(-k_a C_0 t)$ as expected for an ideal catalytic reaction. If $A_0 > D_0$ the overall kinetics of the reaction is that of an autocatalytic reaction. In fact, for $k_a = k_b$ Eq. 8 looks very similar to the logistic equation we obtain when solving a simple autocatalytic reaction. The different limiting cases for the amount of product P_∞ for $t \to \infty$ follow from the discussion above if we make the substitution $D_0 \to D_0 - A_0$.

A linear classifier

Based on the fixed gain amplifier systems explained above we can now build a linear classifier that implements a function

$$\sum_i \alpha_i [m_i] = T. \tag{9}$$

Here α_i are the weights, $[m_i]$ the concentrations of the molecular species m_i and K is the threshold. A molecular implementation of this function thus requires that an initial concentration of m_i results in a concentration $\alpha_i[m_i]$ of some signal molecules that can be compared to each other and to the concentration K of a threshold molecule.

An element of the sum with a positive weight α_i is implemented as a catalytic reaction with a fixed gain α_i. An input m_i at initial concentration $[m_i]_0$ results in a final concentration $\alpha_i[m_i]_0$ of an output strand AP of unrelated sequence. Importantly, the output strand is the same for all reactions with a positive α_i. Similarly, every reaction with a negative α_i is implemented as a catalytic reaction with a (positive) gain $|\alpha_i|$ but a different output strand BP.

In principle, we could use reporters with two different colors to independently read out the the positive and negative output strands AP and BP. Using fluorescence calibration curves, we could then compute the respective concentrations as well as the difference between them and compare the result to the threshold value T. However, such an approach would still require considerable intervention form an experimentalist meaning that only part of the computation is actually implemented as molecular computation.

To embed the comparison of the concentrations of AP and BP in the DNA molecules themselves, we use the annihilator gate design presented in Ref [13] (see also Fig. 9-5). In this design, each of AP and BP bind to annihilator gate G reversibly, but the combination of the two irreversibly binds to G, removing both from solution (Fig. 9-5). In an excess of annihilator gate G, only one of AP and BP will be present in solution at significant concentration. G is present in solution from the beginning of the beginning of the reaction, and serves to dynamically reduce the concentrations of both AP and BP. Note that a similar mutual annihilation reaction could also be implemented using the mechanism for

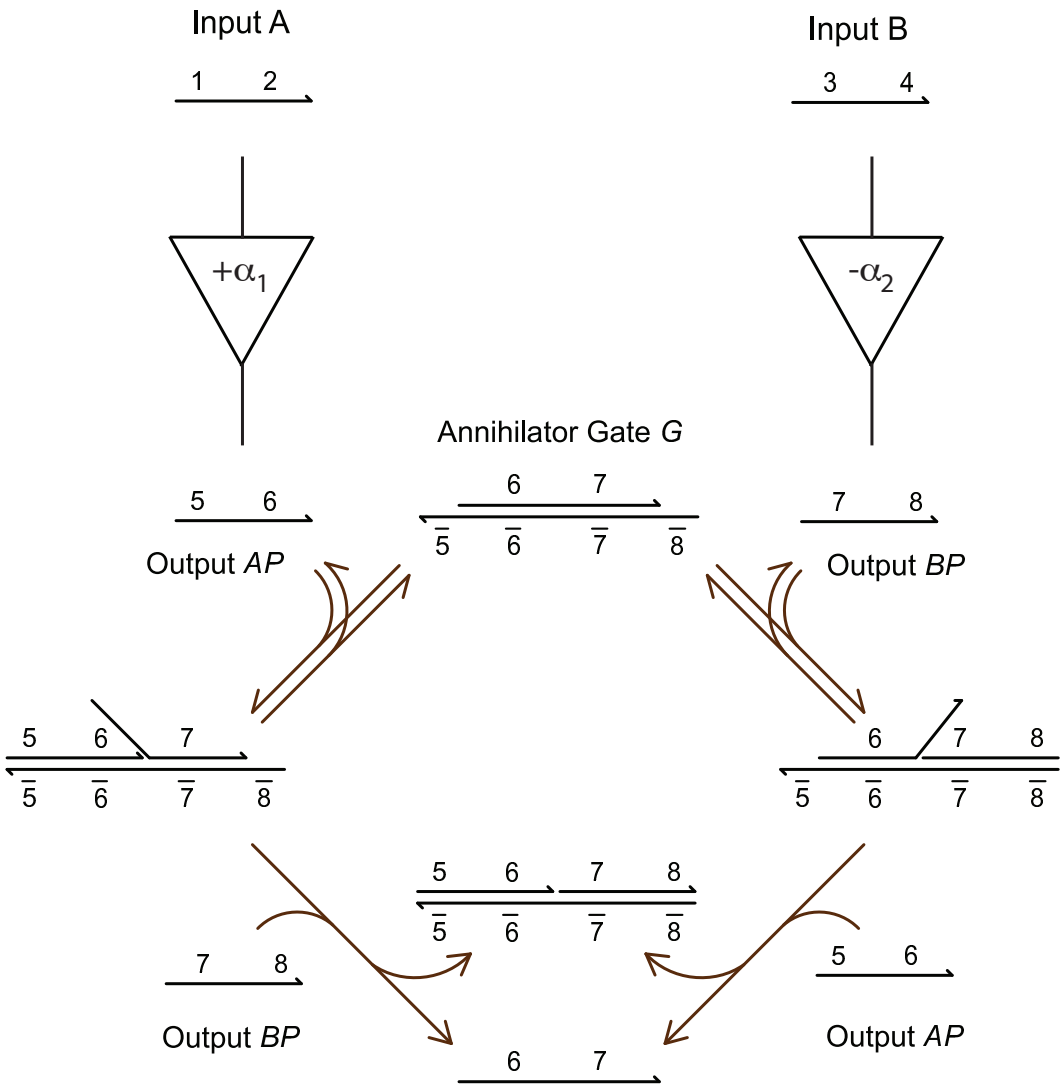


FIG. 9-5: Implementing negative gain. We implement negative gain by having all inputs with positive gains catalytically produce one product AP, and all inputs with negative gains catalytically produce another product of independent sequence, BP. The products AP and BP stoichiometrically neutralize one another via the annihilator gate AG [13]. Excess AP at the end of the reaction denotes that the density classification expression evaluated to positive, while excess BP denotes the expression evaluated to negative.

implementing arbitrary bimolecular reactions explained in Ref. [58].

So far we have shown how to implement arbitrary positive and negative gains and how to perform molecular-level comparison of the concentrations of the resulting reporter strands AP or BP. This would be sufficient to implement a classifier with $T = 0$. To implement a non-zero value for the threshold T we simply add T units of AP or BP depending on the sign of T. In this way we can implement a molecular classifier with arbitrary values for α_i and T on the molecular level.

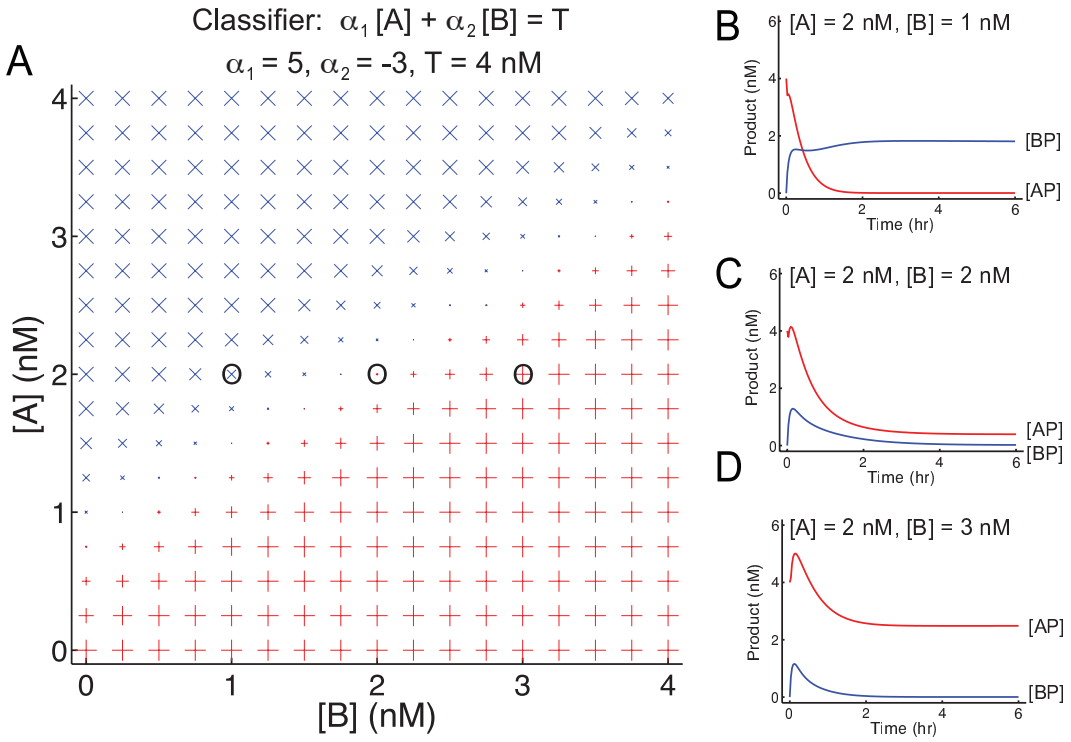


FIG. 9-6: DNA classifier. **(A)** Summary plot of the concentrations of AP and BP at the end of 6 hours of simulated reaction for various initial concentrations of A and B. Size of crosses denote the final concentration of AP; size of pluses denote final concentration of BP. **(B), (C), (D)** Sample concentration traces for AP and BP.

Fig. 9-6 shows an example of a simulation of a simple two-input linear classifier. The simulations use a realistic model for the underlying DNA reactions. Fig. 9-6A shows the expected final signal (i.e. the excess amount of AP or BP) for a variety of "samples." Each sample is characterized by a pair (A, B) of the two molecules of interest. Note that without further amplification of the final output (either AP or BP) the signal linearly increases with the distance from the threshold line.

Conclusions

Here we have proposed a DNA implementation of a fixed gain amplifier and of linear classifier circuits. The fixed gain amplifier combines a DNA catalytic amplifier with a threshold element or an autocatalytic reaction in order to obtain arbitrary gain that can be lower or higher than the intrinsic gain of the DNA catalyst. Classifier circuits similar to the one proposed here can potentially be used for the embedded analysis of RNA expression

levels in complex mixtures. Such classification circuits could find applications in point-of-care diagnostics or could even be used to analyze gene expression in living cells.

To apply the presented linear classifier circuit to actual cell state classification, however, the classifier must be able to deal with RNA input concentrations that are often low and can vary by orders of magnitude. While in theory the methods presented should be able to allow indefinitely high values of α, the precise control of large values of α will be difficult in practice, because the intrinsic turnover set by strand purities will not be known to great accuracy. Additionally, achieving high turnover will be slow, because each turnover requires a fixed amount of time for reaction.

Multi-stage fixed turnover amplifiers can be used to combat the aforementioned difficulties. That is, the products AP and BP can be themselves amplified by another fixed gain amplifier, and the gains of the two systems will be multiplied. Achieving high turnovers with a 2-stage system will also be quadratically faster. For extremely high turnovers, even more stages of fixed amplification can be cascaded.

There are a variety of alternatives to the specific implementation proposed here. In particular, the chemical reaction systems networks of Ref. [58] can be used to implement the reactions described here. However, the catalytic system of Ref. [16] is currently the best characterized and also fastest catalytic amplifier available which is why we chose to use this system for our design.

The reactions and mechanisms used to construct the linear classifier have either been demonstrated or are similar enough to well-understood reactions that they are expected to experimentally function as designed. All simulation results shown include modeling of relevant intermediate species and side reactions; similar modeling has been able to quantitatively predict the kinetics of similar DNA constructions [5, 19]. Thus, we are optimistic that we can experimentally demonstrate the density classifier circuit *in vitro* in the near future.

Acknowledgments. DYZ is supported by the Fannie and John Hertz Foundation. GS is supported by a Career Award at the Scientific Interface from the Burroughs Wellcome

Fund and an NSF CAREER award.

---

[1] V. A. Bloomfield, D. M. Crothers, I. Tinoco Jr. *Nucleic Acids: Structures, Properties, and Functions* **2000,** University Science Books, Sausalito, CA.

[2] J. SantaLucia, D. Hicks, *Annu. Rev. Biophys. Biomol. Struct.* **33,** 415 (2004).

[3] R. M. Dirks, J. S. Bois, J. M. Schaeffer, E. Winfree, N. A. Pierce, *SIAM Rev.* **49,** 65 (2007).

[4] Zuker, M. *Nucleic Acids Res.* **31,** 3406 (2003).

[5] D. Y. Zhang, E. Winfree, *J. Am. Chem. Soc.* **131,** 17303 (2009).

[6] J. Bath, A. J. Turberfield, *Nat. Nanotech.* **2,** 275 (2007).

[7] H. Masu, A. Narita, T. Tokunaga, M. Ohashi, Y. Aoyana, S. Sando, *Angew. Chemie Int. Ed.* **48,** 9481 (2009).

[8] J. Lu *et al. Nature* **435,** 834 (2005).

[9] N. Rosenfeld *et al. Nat. Biotech.* **26,** 462 (2008).

[10] G. Seelig, D. Soloveichik, D. Y. Zhang, E. Winfree, *Science* **314**, 1585 (2006).

[11] M. Hagiya, S. Yaegashi, K. Takahashi, *Nanotechnology: Science and Computation*, 293-308 (2006).

[12] B. M. Frezza, S. L. Cockroft, M. R. Ghadiri, *J. Am. Chem. Soc*, **129**, 14875 (2007)

[13] D. Y. Zhang, "Cooperative DNA strand displacement for DNA quantitation, detection, and logic." (submitted, 2010).

[14] Y. Benenson, B. Gil, U. Ben-Dor, R. Adar, E. Shapiro, *Nature* **429**, 423 (2004).

[15] Z. Xie, S. J. Liu, L. Bleris, Y. Benenson, *Nuc. Acids Res.* (2010, DOI: 10.1093/nar/gkq117).

[16] D. Y. Zhang, A. J. Turberfield, B. Yurke, E. Winfree, *Science* **318**, 1121 (2007).

[17] A. J. Turberfield, J. C. Mitchell, B. Yurke, A. P. Mills, M. I. Blakey, F. C. Simmel, *Phys. Rev. Lett.* **90**, 118102 (2003).

[18] D. Y. Zhang, E. Winfree, *J. Am. Chem. Soc.* **130,** 13921 (2008).

[19] D. Y. Zhang, E. Winfree, *Nuc. Acid Res.* (2010, pre-published online DOI:10.1093/nar/gkq088).

[20] G. Seelig, B. Yurke, E. Winfree, *J. Am. Chem. Soc.* **128,** 12211 (2006).

[21] J. S. Bois, S. Venkataraman, H. M. T. Choi, A. J. Spakowitz, Z. G. Wang, N. A. Pierce, *Nuc. Acid Res.* **33,** 4090 (2005).

[22] S. J. Green, D. Lubrich, A. J. Turberfield, *Biophysical Journal* **91,** 2966 (2006).

[23] B. Yurke, A. J. Turberfield, A. P. Mills, F. C. Simmel, J. L. Neumann, *Nature* **406**, 605 (2000).

[24] R. M. Dirks, N. A. Pierce, *Proc. Nat. Acad. Sci.* **101**, 15275 (2004).

[25] P. Yin, H. M. T. Choi, C. R. Calvert, N. A. Pierce, *Nature* **451**, 318 (2008).

[26] T. Omabegho, R. Sha, N. C. Seeman, *Science* **324,** 67 (2009).

[27] Y. Gao, L. K. Wolf, R. M. Georgiadis, *Nuc. Acids Res.* **34,** 3370 (2006).

[28] W. Sun, C. Mao, F. Liu, N. C. Seeman, *J. Mol. Biol.* **282,** 59 (1998).

[29] R. M. Dirks, M. Lin, E. Winfree, N. A. Pierce, *Nucleic Acids Res.* **32,** 1392 (2004).

[30] J. Seifferf, A. Huhle, *J. Biomol. Struct. Dyn.* **25,** 453 (2008).

[31] F. Tanaka, A. Kameda, M. Yamamoto, A. Ohuchi, *Nuc. Acids Res.* **33,** 903 (2005).

[32] D. Tulpan, M. Andronescu, S. B. Chang, M. R. Shortreed, A. Condon, H. H. Hoos, L. M. Smith, *Nuc. Acids Res.* **33,** 4951 (2005).

[33] J. Sager, D. Stefanovic, *Lecture Notes of Computer Science* **3892,** 275-289 (2006).

[34] N. C. Seeman, *J. Biomol. Struct. Dyn.* **8,** 573-581 (1990).

[35] S. Cao, S. Chen, *Nuc. Acids Res.* **34,** 2634 (2006).

[36] A. Xayaphoummine, T. Bucher, H. Isambert, *Nuc. Acids Res.* **33,** W605 (2005).

[37] R. M. Dirks, N. A. Pierce, *J. Comput. Chem.* **25,** 1295 (2004).

[38] E. Protozanova, P. Yakovchuk, M. D. Frank-Kamenetskii, *J. Mol. Biol.* **342**, 775 (2004).

[39] D. V. Pyshnyi, E. M. Ivanova, *Russian Chemical Bulletin* **51**, 1145 (2002).

[40] V. A. Vasiliskov, D. V. Prokopenko, A. D. Mirzabekov, *Nuc. Acid Res.* **29**, 2303 (2001).

[41] D. V. Pyshnyi, E. M. Ivanova, *Nucleosides, Nucleotides, and Nucleic Acids* **23**, 1057 (2004).

[42] I. G. Panyutin, P. Hsieh, *J. Mol. Biol.* **230,** 413 (1993).

[43] D. Sen, W. Gilbert, *Methods in enzymology* **211,** 191 (1992).

[44] E. M. Southern, S. C. Casegreen, J. K. Elder, M. Johnson, K. U. Mir, L. Wang, J. C. Williams, *Nuc. Acids Res.* **22,** 1368 (1994).

[45] Mir, K. U. A restricted genetic alphabet for DNA computing. *DNA based computers II. DIMACS* **44,** 243-246 (1998).

[46] L. Qian, E. Winfree, *Lecture Notes of Computer Science* **5347,** 70 (2009).

[47] C. Mao, W. Sun, Z. Shen, N. C. Seeman, *Nature* **397,** 144 (1999).

[48] L. P. Reynaldo, A. V. Vologodskii, B. P. Neri, V. I. Lyamichev, *J. Mol. Bio.* **297**, 511 (2000).

[49] B. Yurke, A. P. Mills, *Genet. Prog. Evol. Mach.* **4**, 111 (2003).

[50] M. Zuker, D. H. Mathews, D. H. Turner, "Algorithms and Thermodynamics for RNA Secondary Structure Prediction: A Practical Guide In RNA Biochemistry and Biotechnology," in J. Barciszewski and B.F.C. Clark, eds., NATO ASI Series, Kluwer Academic Publishers, (1999).

[51] R. A. Dimitrov, M. Zuker, *Biophys. J.* **87,** 215 (2004).

[52] J. Kim, K. S. White, E. Winfree, *Mol. Syst. Biol.* **2,** 68 (2006).

[53] W. U. Dittmer, F. C. Simmel *Nano Lett.* **4,** 689 (2004).

[54] Stojanovic, M. N., Semova, S., Kolpashchikov, D., Macdonald, J., Morgan, C. & Stefanovic, D. *J. Am. Chem. Soc.* **127,** 6914-6915 (2005).

[55] R. Pei, S. K. Taylor, D. Stefanovic, S. Rudchenko, T. E. Mitchell, M. N. Stojanovic, *J. Am. Chem. Soc.* **128,** 12693 (2006).

[56] K. Lund, A. Manzo, N. Dabby, N. Michelotti, A. Johnson-Buck, J. Nangreave, S. Taylor, R. Pei, M.

N. Stojanovic, N. Walter, E. Winfree, H. Yan, *Nature* (in press, 2010).

[57] M. D. Frank-Kamenetskii, S. M. Mirkin, *Annu. Rev. Biochem.* **64,** 65 (1995).

[58] D. Soloveichik, G. Seelig, E. Winfree, *Proc. Nat. Acad. Sci.* (2010, pre-published online DOI:10.1073/pnas.0909380107).

[59] A. Phillips, L. Cardelli, *Journal of the Royal Society Interface* **6,** S419 (2009).

Chapter 10: Towards Self Replication

In Chapter 4, I experimentally demonstrated an autocatalytic reaction, in which the autocatalyst molecule catalyzed the production of identical copies of itself. I was (and still am) hesitant to call it alive, or even a "self-replicator," because there isn't any fitness advantage for the autocatalyst molecule to release a copy of itself, rather than some other molecule. In a sense, because it was not a templated reaction, there was necessarily a lack of information flow across generations, and thus that system is incapable of Darwinian evolution, and not "alive" by my standards.

The work below shows a Rube Goldberg-ish design for a replicating super-structure composed of DNA that, in principle, is capable of Darwinian evolution. Of course, this design is way too complicated to test experimentally; even if the entire process works, "sequencing" the replication product is no small task! Personally, I think building a (much more simplified) form of evolving life out of DNA may not be too far off in the horizon– maybe another four years, maximum, if an industrious graduate student were to put his mind to it. But it's not clear doing so would actually uncover any deep scientific truths or enable any practical technologies... after all, everyone who works in this field presumably understands and believes in evolution.

Bernard Yurke and I came up with the initial idea for this work at Tony Roma's while feasting on ribs in 2003, marking this as my earliest work in the field of DNA computation and engineering. For me, this was an exercise in creativity that provided intangible benefits

in the form of intuition on the behavior of molecules at the nanoscale. For example, the cyclization described in the paper became an often recurring theme and concern in my later studies of DNA.

To any young graduate student who is confused about what to do experimentally, I'd highly recommend doing a similar "thought experiment;" it is from these that you develop your critical thinking skills. Come up with an outlandish plan, and have your adviser poke a million holes in it. Then learn the reasons why your original plan didn't work, and use that to formulate a project that works.

Parts of this work were presented at the 10th International DNA Computing conference in Milan, Italy. This work was published in full as:

Zhang, David Yu; Yurke, Bernard. "A Superstructure-Based Self Replicator Without Product Inhibition." *Natural Computing* 5(2): 183-202 (2006).

A monomer structure based on a hairpin loop is described that can be linked via short oligonucleotide sequences (linkers) to form polymers. Independence of linked monomers allow for exponential complexity of the polymer structure. A method is described wherein the polymer structure can be replicated semi-conservatively with fidelity, given a source of monomer structures and linkers. Furthermore, the separation of the product from the parent allows for exponential amplification. These steps are achieved by secondary structure constraints and toehold-mediated strand displacement, and occur in the absence of enzymes. The parallel polymerization allows for replication to be achieved in O(log N) time, as opposed to O(N) from a processive process.

Keywords: DNA motors, DNA based polymers, self-replication

Introduction

The ability to reproduce is one of the defining characteristics of life. This ability arises from a process by which DNA serves as its own template to make copies of itself [1]. There

is considerable interest in the construction of synthetic systems that are also capable of template-directed replication [2]. For our purposes synthetic systems will denote systems that do not take advantage of biologically derived and evolutionarily advanced enzymes, such as polymerases, to achieve replication of nucleic acids. Much of the interest in synthetic replication systems comes from the insights they may provide on how life originated or functions [3, 4]. There is interest also from the point of view of what template replication may have to offer for manufacturing. Replication by living organisms is an exponential growth process. In a manufacturing setting this kind of replication would allow for easy scale-up of the volume of units produced. Another feature exhibited by living organisms is that the replication process is not perfect. This allows for Darwinian evolution. The power of directed evolution has been exploited in research and manufacturing settings [5]. Error-prone synthetic replicators may also allow for the directed evolution of useful products.

A number of synthetic chemical systems have been constructed which exhibit template-directed replication [6] or self-replication in oligonucleotide-based systems [7]-[15], peptide-based systems [16]-[20], and other chemical systems [21]-[27]. A characteristic problem of these systems is that the product remains bound to the template or competes with monomers for binding with the template. Such systems exhibit sublinear parabolic $O(\sqrt{N})$ growth rather than exponential growth. The replication process tends to stall as the concentration of product increases.

Exponential growth is a prerequisite for selection in the Darwinian sense [28, 29]. The sub-exponential growth exhibited by these systems is thus not conducive to Darwinian evolution. Recently a self-replicating system based on a ligase ribozyme has been constructed [30] that, at least, at early times exhibits exponential growth.

Recently exponential growth has also been demonstrated for a system employing a stepwise 'feeding' procedure and immobilization of the product on a support [31]. In this system a template is immobilized on a solid support. Fragments of the product that bind to the template are introduced. The fragments are then ligated together to produce the product. Next, a denaturing step releases the product which is then immobilized on fresh solid support. The problem of the binding of a template with its product is thus overcome by

immobilization.

Various DNA-based nanostructures [32]-[38] and nanomachines [39]-[46] have been constructed. For a recent review see [38]. This suggests that DNA may be a suitable medium for the construction of synthetic replicators. Most of the DNA-based nanomachines that have been constructed [40]-[46] to date use the energy of hybridization to activate the machines. In addition, strand displacement through competitive binding mediated by toeholds is generally used to return the machine to its initial state. The operation of these machines involves stepwise feeding and, hence, are clocked in the sense that an external operator controls when the machine advances to its next state through the addition of the appropriate DNA strand to the solution containing the machines.

Here we present a design for a clocked replicator utilizing the energy of hybridization to pull the template-replicated product from the template, to allow for exponential growth. Toehold-mediated strand displacement is used to clear the spent replication machinery from the template and the product to prepare for the next round of template replication. The replication process proceeds in a clocked manner in which the replication process is sequenced through a series of steps via the addition of DNA strands in an appropriate sequence. In the discussion that follows, it will be assumed that the DNA strands are added in slight excess to facilitate the quick completion of reactions. The unreacted excess in each step does not influence the outcome of the procedure, since they are consumed by further excesses in the next stage.

It is argued that the replication scheme we propose is not stalled by product-template binding and can exhibit exponential growth in the presence of abundant monomer concentration. If the template-directed replication process is made a bit faulty this system could, in principle, be subjected to directed evolution.

The Superstructure Polymer

The superstructure polymer (*superduplex*) to be replicated is composed of two *superstrands* of polymers S_1 and S_2, much as duplex DNA is composed of two strands. Each superstrand in turn is composed of serially linked monomer *superbases* (see Fig. 10-1).

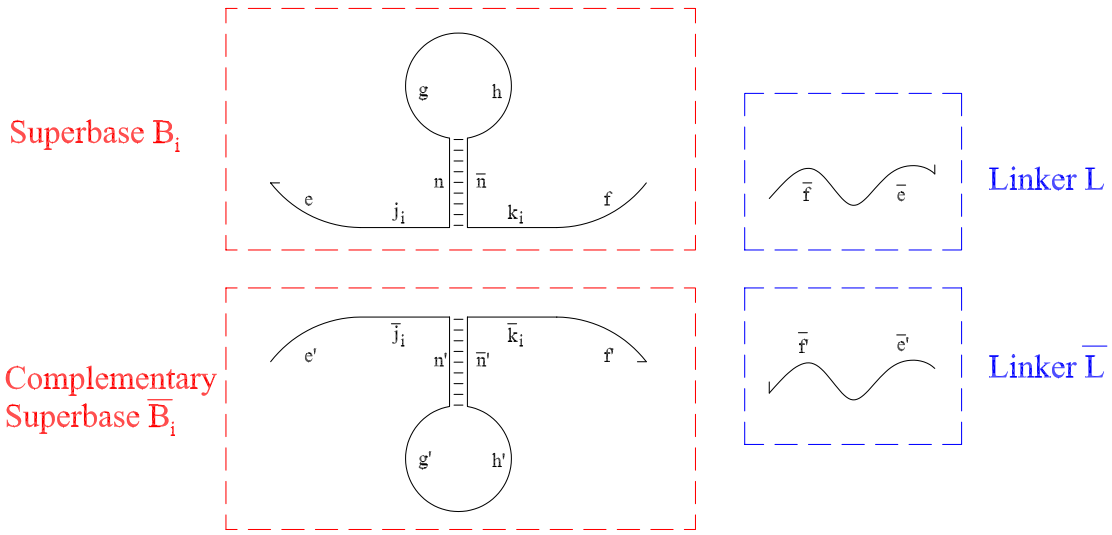


FIG. 10-1: The monomer superbases and the linker strands.

However, unlike DNA, the two superstrands are not identical.

Each different superbase B_i on superstrand S_1 is a single DNA oligonucleotide strand with a hairpin structure. The information-containing regions of the superbase B_i are the domains j_i and k_i. The complementary regions n and $\bar{n}$ form the stem of the hairpin that defines the structure of the free monomer superbase. Domains g and h form the loop of the hairpin, and are used for separating the two parental superstrands during replication. The dangling domains e and f serve to link to adjacent superbases to form the superstrand polymer (see Fig. 10-2).

The i subscript of the superbase B_i refers to its identity, not its position in the superstrand. For example, B_1 could represent the superstructure analog of the DNA nucleotide base thymine, while B_2 represents cytosine.

Each superbase B_i on superstrand S_1 has a unique complementary superbase $\bar{B}_i$ on

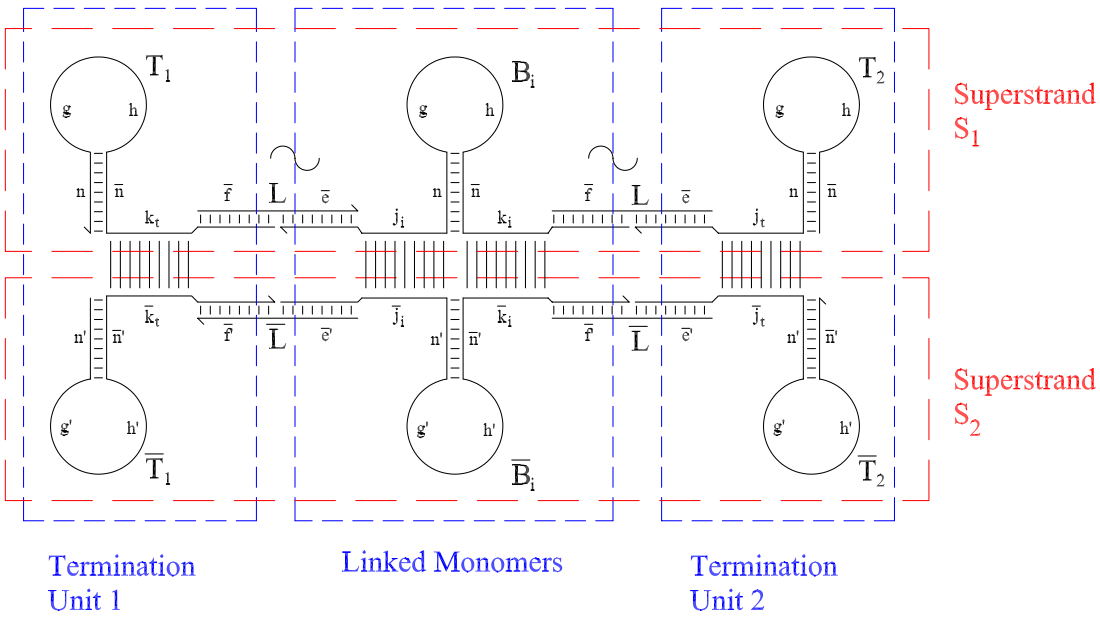


FIG. 10-2: The formed superduplex, composed of two superstrands.

superstrand S_2 to which it binds perfectly. The complementary superbase $\bar{B}_i$ is very similar to B_i, possessing also a hairpin structure, two information domains (which are necessarily complementary to the corresponding ones on B_i), and two linking domains.

The domains e, f, g, h, n, and $\bar{n}$ can be identical for all bases B_i, but are necessarily different for the two superstrands. Arbitrarily, we name the domains on the S_1 strand e, f, g, h, n, and $\bar{n}$, and the corresponding domains on the S_2 strand e', f', g', h', n', $\bar{n}'$. The n' and n domains (and other analogous pairs) are similar only in function and are completely unrelated in sequence. Specifically, they are not complements; the four-way junctions depicted in Fig. 10-2 are stable. The complement of a domain p is denoted as $\bar{p}$.

The linker strands L and L' serve as the analog of the phosphate bonds which connect consecutive bases of nucleotides in DNA, connecting superbases to form superstrands. L is composed of the domains $\bar{e}$ and $\bar{f}$, complements to the similarly-named domains on the

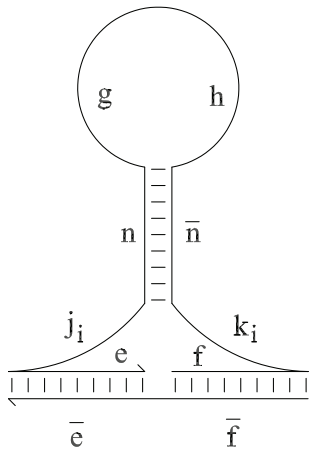


FIG. 10-3: A cyclized monomer.

superbases of S_1. Similarly, L' is composed of the domains $\bar{e}'$ and $\bar{f}'$. Connection of the superbases is effected by the simultaneous binding of the linker strands to two different superbases, as shown in Fig. 10-2.

Fig. 10-2 also show four special monomers, T_1, $\bar{T}_1$, T_2 and $\bar{T}_2$, which serve as chain terminators. They are similar to the standard superbase B_i, except they have only one linking domain (e or f), and one information domain (j or k). These special terminators are necessary to prevent the random saltation of monomers onto the ends of the parental superstrands during the replication process.

With the given structure of the superbases and the linkers, it is worthwhile to point out that monomers may cyclize as in Fig. 10-3. This feature of the system will actually be useful during the replication process, in inactivating excess monomers.

Binding of the two superstrands S_1 and S_2 is specific, due to the recognition domains j, k, $\bar{j}$, and $\bar{k}$. However, just as with DNA, two superstrands that are mostly complementary can associate despite the presence of a few mismatches. Also, just as with DNA, the binding of a single superbase with its complement should not be irreversibly strong. Thus, information domains j and k should be short, on the order of five bases.

The Replication Process

An overview flowchart of the replication process is shown in Fig. 10-4. As can be seen, the process can be compartmentalized into five distinct steps, each triggered by the manual

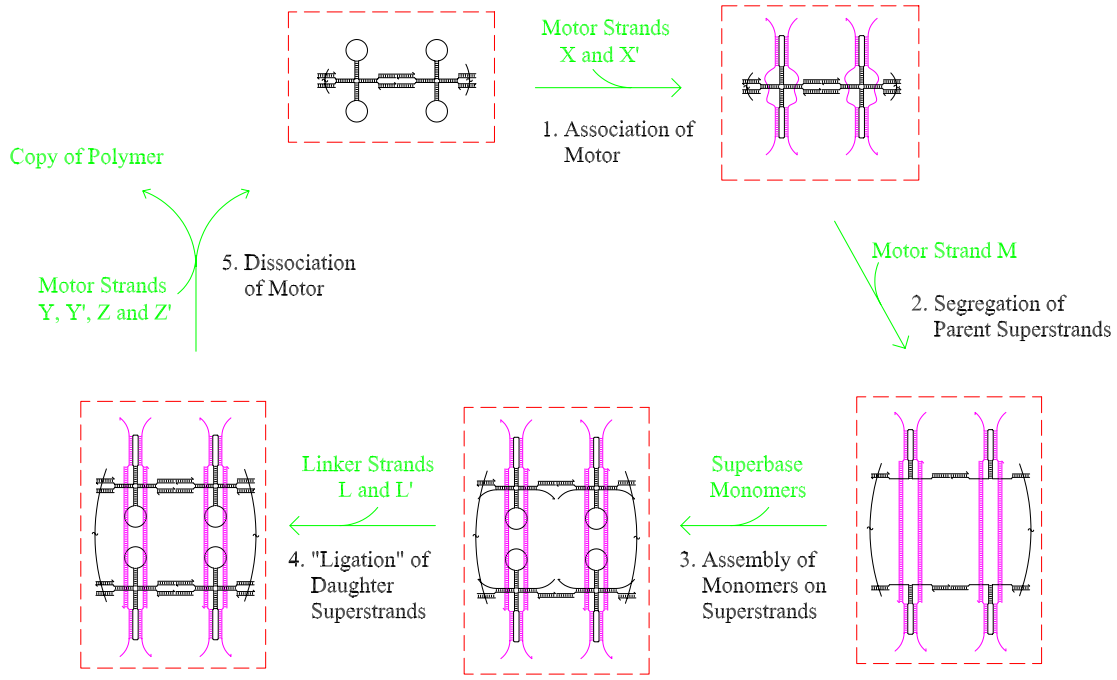


FIG. 10-4: Flowchart of the replication process.

addition of one set of DNA strands. For the purposes of this section, we will only consider what happens in the proper function of the replication process, under perfect stoichiometric conditions. Considerations for excess reagents will be made in the Discussions section.

Replication involves the use of a "motor" apparatus, the components of which are depicted in Fig. 10-5. There are seven different oligonucleotide strands involved, labelled X, X', Y, Y', Z, Z', and M. The strands X and X' associate with the double-superstranded polymer that is to be replicated first, and then the two parental strands are separated physically with the use of effector strands M. Strands Y, Y', Z, and Z' serve to detach the X and X' strands after the replication process is complete so that a new round of replication may begin. At the end of a replication cycle, each instance of X is bound to Y, Z, and M, and the resulting $XYZM$ complex is a fully doubled-stranded inert waste product. Similarly, $X'Y'Z'M$ is also formed, and also inert.

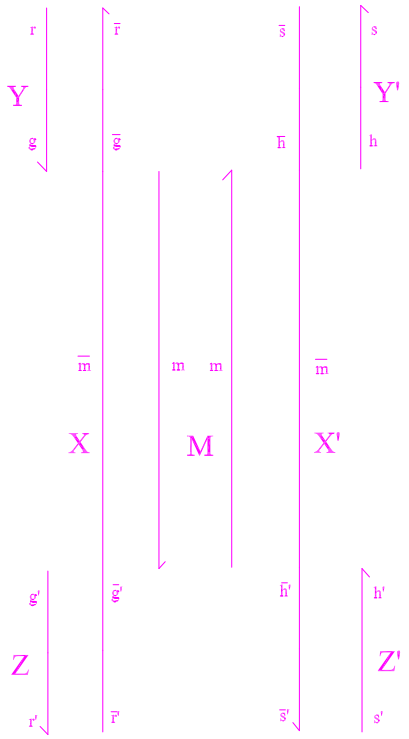


FIG. 10-5: The motor apparatus that effects the separation of the two parent superstrands.

The hybridization energy of the m and $\bar{m}$ domains on the motor strands provide the driving force that serves to pull the two parental strands apart, and necessarily must be long, at very minimum longer than the sum of the lengths of the j and k information domains of the superbases.

Step 1: Association of the Motor Apparatus

The first step, the association of the motor apparatus, is triggered by the external addition of the X and X' strands. They associate with the superduplex as in Fig. 10-6.

The hybridization of the X and X' domains to their respective complements on the loop region are relatively straightforward. The loop region itself is geometrically constrained, but both X and X' are single-stranded, and can wind around the loop to form the double-helix of the g, g', h, and h' domains.

This step of the replication cycle is the one most prone to error, since each motor strand

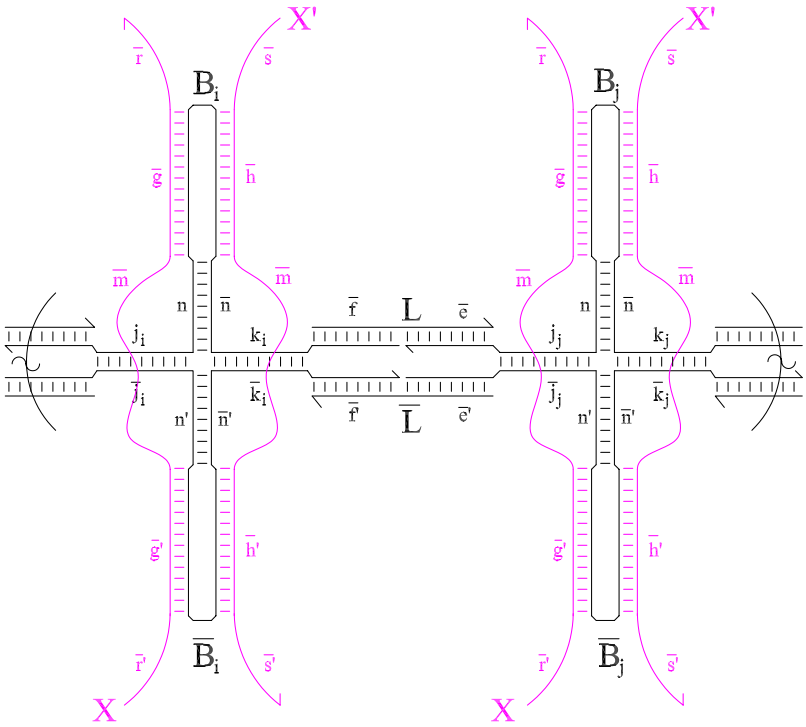


FIG. 10-6: The superduplex after the addition of the motor strands X and X'.

X (or X') must simultaneously bind two different domains relatively far from each other. Fig. 10-6 shows the correct, and only one of many possible, ways for the X and X' strands to associate with the superstrands.

One possibility of error is that an X strand with its 3' end bound to the $\bar{g}$ of B_i might bind to $\bar{g'}$ of $\bar{B}_j$, rather than the corresponding domain on $\bar{B}_i$. This type of error can be arbitrarily disfavored by adjusting the relative lengths of the m, e, and f domains.

Another possibility of error is that of dimerization and aggregation. This occurs when an X strand simultaneously binds to both the g domain of an S_1 superstrand and the g' domain of an S_2 superstrand on a different superduplex. If this occurs, then the two superduplexes would be tethered together, causing unwanted reactions in later steps. We propose no universal solution to this problem, but note that this is unlikely to happen when the concentration of the superduplexes is low. As the concentration of the superduplex

increases through repeated rounds of replication, this effect may set the upper limit on production.

The loop formed by the single-stranded m domain can be either in front of or behind the linker duplex. The two configurations are energetically equal, and thus approximately half of the loops will be behind the duplex. The position of the m loops does not matter for the replication process.

Step 2: Segregation of the Parent Superstrands

Upon the addition of the M strands, the m domains bind to the $\bar{m}$ domains on X and X' (Fig. 10-7). The first few base pairs form normally. However, since duplex DNA adopts a double-stranded structure with substantial persistence length (on the order of 40nm, or 120 base pairs), as more base pairs form along the m domain, superstrands S_1 and S_2 are forced apart.

The free energy gained by the hybridization of the base pairs of m to $\bar{m}$ favor the segregation of the parent strands, while the already formed base pairs of the information domains j and k oppose this. Due to the geometry of the system, every base pair formed in the m domain by X and X' will break one base pair of either j or k domains. Thus, this stage is an unbiased random walk, just as normal branch migration is.

The segregation of the parent superstrands is favored by making the m domain much longer than the j and k domains, so that j and k are kept physically far away from $\bar{j}$ and $\bar{k}$, and no reverse reaction is possible. As mentioned previously, j and k should be on the order of five bases. The length of domain m is recommended to be on the order of 30 bases; this choice is in consideration of making the critical Euler buckling force larger than the force required to pull a base pair apart, so that separation is favored over the motor domain $m\bar{m}$ developing a kinked structure.

The actual hybridization of the M strand to the X and X' strands is a simple DNA hybridization process that is unlikely to encounter errors, assuming that the X and X' strands had associated with the superduplex correctly in the previous step.

Step 3: Assembly of Monomers on Parental Superstrands

Following the segregation of the two parent superstrands, superbase monomers of both

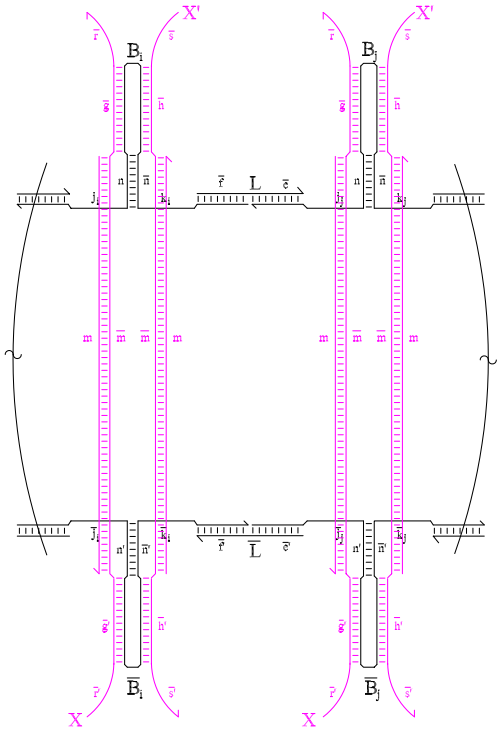


FIG. 10-7: Pulling apart the two superstrands.

B and $\bar{B}$ types are added. These will seek out their complements on the separated parent superstrands, and bind transiently by the information domains j and k, as shown in Fig. 10-8.

The chance of error on this step is low, because the two domains by which the monomers bind (j and k) are both weak. The large loop entropy of bridging over a distance to bind two different superbases will destabilize the thermodynamics enough to cause the monomer superbase to spontaneously dissociate from its erroneous binding.

Step 4: "Ligation" of Daughter Superstrands

Since the appropriate sequence of monomers has already been assembled on both parental superstrands, all that remains in finishing the replication step is to ligate them into a new superstrand. Ligation, as used in this paper, does not refer to the formation of covalent bonds using the biologically derived enzyme ligase. Rather, we use the linker strands L and

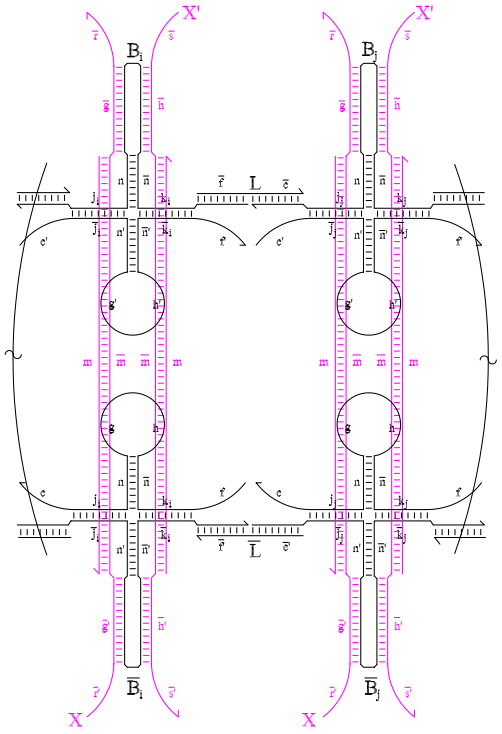


FIG. 10-8: Parental superstrands direct the binding of the appropriate monomers for forming daughter superstrands.

L' to act as permanent splints, which join connected superbases. See Fig 10-9.

We do not wish for the ligated superstrands to dissociate spontaneously within the lifetime of the reaction; therefore the linking domains e and f should be reasonably long, on the order of 20 base pairs each.

The process of "ligating" the monomers is a straight-forward hybridization reaction. The f and e domains of adjacent monomers are held to fairly close proximity, so the chance of incorrect binding by the linker strands is low.

Step 5: Dissociation of the Motor Apparatus

The process of replication is complete by this point; we started with two parental superstrands, and we now have two additional daughter superstrands. Just as DNA, replication was semi-conservative; each of the superduplexes contain one parental strand and one daughter strand.

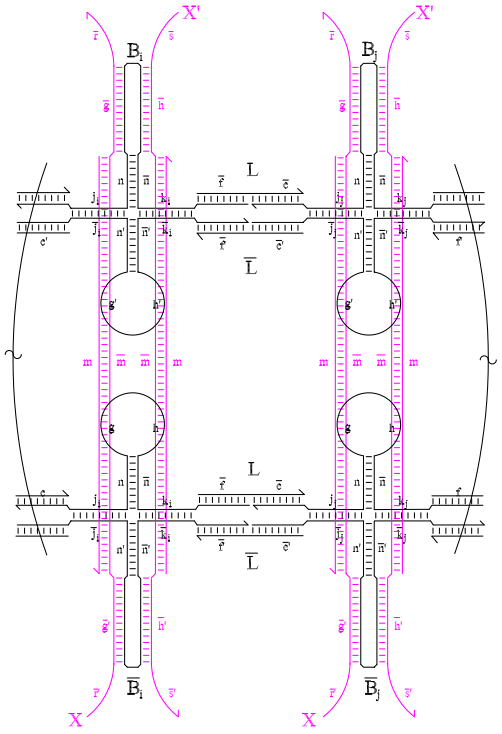


FIG. 10-9: Addition of L and L' linker strands ligate the assembled monomers bound to the two parental superstrands, forming daughter superstrands.

All that remains to be done in this step is to remove the motor apparatus and separate the two superduplexes to allow another round of replication. This is achieved with the addition of the four motor removal strands Y, Y', Z, and Z'. All four function similarly, so only the process for Y will be described.

First, a molecule Y binds to every instance X via the toehold domain r. This allows the Y strand to be localized near the X strand, greatly increasing the effective concentration. A simple branch migration process occurs, until Y is bound to X by the entirety of the g domain as well, at which point strand X is no longer constrained to be near the loop region of B_i, and no reverse reaction is possible.

Upon the completion of this step, the system is restored to its original state, shown in Fig. 10-10. There are a number of inert $XYZM$ and $X'Y'Z'M$ duplexes, and twice the

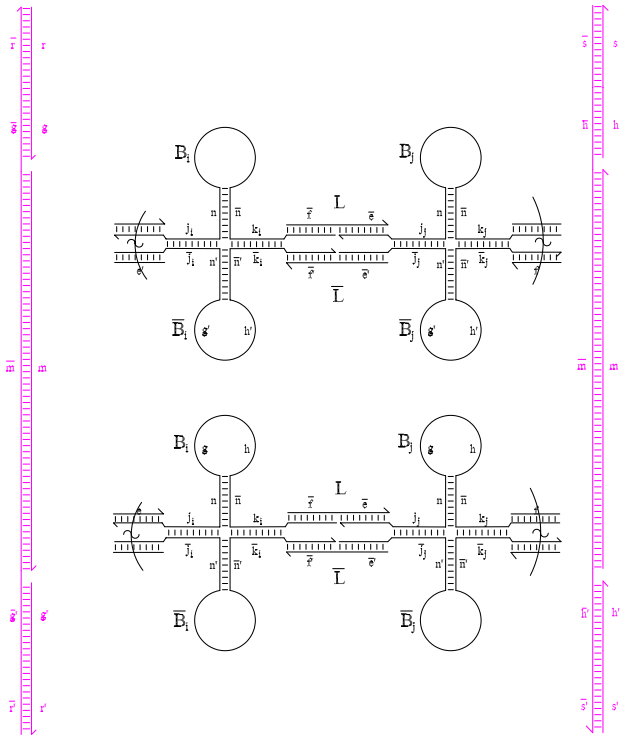


FIG. 10-10: Addition of Z and Z' strands cause the motor to detach from the superstructure, releasing two identical daughter superduplexes.

number of superduplex polymers as from the beginning of the replication cycle.

Toehold-mediated branch-migration used in this step is well characterized and extensively tested in the laboratory; thus it is not likely for an error to occur.

Optionally, an excess clean-up step can be inserted before starting another replication cycle. Though probably not necessary for short procedures, it may be needed if the number of cycles is large since the accumulation of excess strands may substantially alter the kinetics and fidelity of the replication. See the Discussion section for details.

Discussion

Stoichiometric Issues

Addition of reagents at every step of the replication process should be in slight excess. This is necessary for two purposes: first, experimentally it is not possible to achieve perfect

stoichiometry. Second, perfect 1:1 stoichiometry would cause most reactions to possess second-order kinetics, rather than first-order kinetics in the case of excesses. Excesses serve to speed up the replication process considerably.

We consider the effects of having excess DNA strands in solution during each of the steps:

Step 1: The strands being added are X and X'. Excess X and X' will float in solution, since no other binding sites for them are available.

Step 2: The strand being added is M. Excess M will bind to the excess X and X' from Step 1, and form XM and $X'M$, with some leftover M.

Step 3: The strands being added are superbase monomers B and $\bar{B}$. XM and $X'M$ formed in excess reactions of Step 2 may bind to either the assembling or the excess monomers on the loop domains g, g', h, or h'. This will have no substantial effect other than possible slowing of the kinetics of the ligation of Step 4 due to steric hindrance. Excess B and M are inert.

Step 4: The strands being added are L and L'. Excess L and L' will cause excess monomers B and $\bar{B}$ from Step 3 to cyclize (see Fig. 10-3), inactivating them from further reaction.

Step 5: The strands being added are Y, Y', Z, and Z'. Excess of Y, Y', Z, and Z' will displace the improperly bound XM and $X'M$ on the daughter strands, from Step 3, and form the inert $XYZM$ and $X'Y'Z'M$.

At the end of a replication cycle, the excesses that remain are M, Y, Y', Z, Z', L, L', cyclic B, and cyclic B'.

Cyclic B and cyclic B', as mentioned before, are inert. L and L' will inactivate some of the B and B' monomers by cyclizing them in Step 3 of the next cycle before they assemble on the superstrands, but is otherwise harmless to the replication process.

Excess Y, Y', Z, and Z' will bind to and deactivate the X and X' strands during Step 2 of the next cycle. If only one end of X or X' is deactivated while the other end binds to a superbase, then addition of motor strand M will fail to segregate the parent superstrands at that base, causing non-disjunction. Excess M will bind to X and X' prematurely in Step

1, and favor the dimerization and aggregation described in Section 3.1.

Thus, we must clean up excesses of Y, Y', Z, Z', and M before the next cycle of replication. This can be done by adding excesses of the $\bar{Y}$, $\bar{Y}'$, $\bar{Z}$, $\bar{Z}'$, and $\bar{M}$ strands (not shown), which are the full complements of the respective strands. Excesses of $\bar{Y}$, $\bar{Y}'$, $\bar{Z}$, $\bar{Z}'$, and $\bar{M}$ will do nothing except inactivate some of their respective complements in the next cycle.

Parallel Assembly

The assembly of monomer superbases onto the parent superstrands is parallel, rather than serial. Polymerization is not processive–the failure of a monomer to bind at a position will lead to two truncated fragments, rather than one. There is no clear indication of whether the system is ready to proceed onto the next step of replication, so fragmentation rate will be higher than that of a system which waits for each base to be assembled before continuing.

The parallel nature of the replication process, however, also allows for faster kinetics. Whereas any serial (processive) replication process requires $\mathrm{O}(N)$ time, where N is the number of superbases, parallel assembly decreases the asymptotic time complexity of one replication cycle to $\mathrm{O}(\log N)$. In the presence of excess monomers, the kinetics of assembly approaches first-order.

This allows for the fast replication of very long sequences of superduplexes.

Conclusions

We have described a DNA superstructure-based polymer capable of information storage and its associated mechanism for replication. It resembles DNA in that it is double-superstranded and that its replication is semi-conservative, but it does not require any enzymes for replication. Our replication design possesses a mechanism for parental superstrand segregation so that product inhibition is not a problem. Thus the system should exhibit amplification exponential in the number of cycles performed. Due to the parallel nature of the assembly process, each cycle of our replication process should take $\mathrm{O}(\log N)$ time, which is superior to the processive biological methods that take $\mathrm{O}(N)$ time.

Acknowledgements

The authors would like to thank Robert Barish and Si-ping Han for proofreading assistance.

[1] Alberts B (2003) DNA replication and recombination. Nature 421: 431-435

[2] Wintner EA, Conn MM, and Rebek J (1994) Studies in molecular replication. Acc. Chen. Res. 27: 198-203

[3] Penrose LS (1959) Self-Reproducing Machines. Scientific American 200(6): 105-114

[4] Orgel LE (1995) Unnatural Selection in Chemical Systems. Acc. Chem. Res. 28: 109-118

[5] Bull JJ and Wichman HA (2001) Applied Evolution. Annu. Rev. Ecol. Syst. 32: 183-271

[6] Johnston WK, Unrau PJ, Lawrence MS, Glasner ME, and Bartel DP (2001) RNA-Catalyzed RNA Polymerization: Accurate and General RNA-Templated Primer Extension. Science 292: 1319-1325

[7] von Kiedrowski G (1986) A self-replicating hexadeoxynucleotide. Angew. Chem. Int. Edn Engl. 25: 932-935

[8] Zielinski WS and Orgel LE (1987) Autocatalytic synthesis of a tetranucleotide analogue. Nature 327: 346-347

[9] von Kiedrowski G, Wlotzka B, Helbing J, Matzen M, and Jordan S (1991) Parabolic growth of a hexadeoxynucleotide analogue bearing a 3'-5'-phosphoamidate link. Angew. Chem. Int. Edn Engl. 30: 423-426

[10] Achilles T and von Kiedrowski G (1993) A self-replicating system from three precursors. Angew. Chem. Int. Edn 32: 1198-1201

[11] Sievers D and von Kiedrowski G (1994) Self-replication of complementary nucleotide-based oligomers. Nature 369: 221-224

[12] Li T and Nicolaou KC (1994) Chemical self-replication of palindromic duplex DNA. Nature 369: 218

[13] Martin B, Micura R, Pitsch S, and Eschenmoser A (1997) Pyranosyl-RNA: further observations on replication. Helv. Chim. Acta 80: 1901-1951

[14] Sievers D and von Kiedrowski G (1998) Self-replication of hexadeoxynucleotide analogues: autocatalysis versus cross-catalysis. Chem. Eur. J. 4: 629-641

[15] Schoneborn H, Bulle J, and von Kiedrowski G (2001) Kinetic monitoring of self-replicating systems through measurement of fluorescence resonance energy transfer. Chembiochem 2: 922-927

[16] Lee DH, Granja JR, Martinez JA, Severin K, and Ghadiri MR (1996) A self-replicating peptide. Nature 382: 525-528

[17] Severin KS, Lee DH, Martinez JA, and Ghadiri MR (1997) Peptide self-replication via template-directed ligation. Chem. Eur. J. 3: 1017-1024

[18] Severin KS, Lee DH, Martinez JA, Vieth M, and Ghadiri MR (1998) Dynamic error correction in

autocatalytic peptide networks. Angew, Chem. Int. Edn Engl. 37: 126-128

[19] Yao S, Ghosh I, Zutshi R, and Chmielewski JA (1998) A self-replicating peptide under ionic control. Angew. Chem. Int. Edn Engl. 37: 478-481

[20] Saghathelian A, Yokobayashi Y, Soltani K, and Ghadiri MR (2001) A chiroselective peptide replicator. Nature 409: 797-801

[21] Tjivikua T, Ballester P, and Rebek JA (1990) A self-replicating system. J. Am. Chem. Soc. 112: 1249-1250

[22] Terfort A and von Kiedrowski G (1992) Self-replication during condensation of 3-aminobenzamidines with 2-formylphenoxyacetic acids. Angew. Chem. Int. Edn Engl. 31: 654-656

[23] Hong JI, Fang Q, Rotello V, and Rebek JA (1992) Competition, cooperation, and mutation improving a synthetic replicator by light irradiation. Science 255: 848-850

[24] Fang Q, Park TK, and Rebek JA (1992) Crossover reactions between synthetic replicators yield active and inactive recombinants. Science 256: 1179-1180

[25] Pieters RJ, Huc I, and Rebek JA (1994) Reciprocal template effect in a replication cycle. Angew. Chem. Int. Edn Engl. 106: 1579-1581

[26] Reinhoudt DN, Rudkevich DM, and de Jong F (1996) Kinetic analysis of the Rebek self-replicating system: Is there a controversy? J. Am. Chem. Soc. 118: 6880-6889

[27] Wang B and Sutherland IO (1997) Self-replication in a Diels-Alder reaction. Chem. Commun. 16: 1495-1496

[28] Szathmary E and Gladkih I (1989) Sub-exponential growth and coexistence of non-enzymatically replicating templates. J. Theor. Biol. 138: 55-58

[29] Wills PR, Kauffman SA, and Stadler BMR (1998) Selection dynamics in autocatalytic systems: Templates replicating through binary ligation. Bull. Math. Bio. 60: 1073-1098

[30] Paul N and Joyce GF (2002) A self-replicating ligase ribozyme. PNAS 99: 12733-12740

[31] Luther A, Brandsch R, and von Kiedrowski G (1998) Surface-promoted replication and exponential amplification of DNA analogues. Nature 396: 245-248

[32] Chen J and Seeman NC (1991) The synthesis from DNA of a molecule with the connectivity of a cube. Nature 350: 631-633

[33] Seeman NC (1998) Nucleic acid nanostructures and topology. Angew. Chem. Int. Edn Engl. 37: 3220-3238

[34] Winfree E, Liu F, Wenzler LA, Seeman NC (1998) Design and self-assembly of two-dimensional DNA crystals. Nature 394: 539-544

[35] Mao C, Sun W, Seeman NC (1999) Designed two-dimensional DNA Holiday junction arrays visualized by atomic force microscopy. J. Am. Chem. Soc. 121: 5437-5443

[36] LaBean TH, Yan H, Kopatsch J, Liu F, Winfree E, Reif JH, Seeman NC (2000) Construction, analysis, ligation, and self-assembly of DNA triple crossover complexes. J. Am. Chem. Soc. 122: 1848-1860

[37] C. Mao, T. H. LaBean, J. H. Reif, and N. C. Seeman, "Logical computation using algorithmic self-assembly of DNA triple-crossover molecules," Nature **407**, 493-496 (2000).

[38] N. C. Seeman, "DNA in a material world," Nature **421**, 427-431 (2003).

[39] C. Mao, W. Sun, Z. Shen, and N. C. Seeman, "A nanomechanical device based on the B-Z transition of DNA," Nature **297**, 144-146 (1999).

[40] B. Yurke, A. J. Turberfield, A. P. Mills, Jr., F. C. Simmel, and J. L. Neumann, "A DNA-fuelled molecular machine made of DNA," Nature **406**, 605-608 (2000).

[41] Simmel FC and Yurke B (2001) Using DNA to construct and power a nanoactuator. Phys. Rev. E 63: 041913

[42] Simmel FC and Yurke B (2002) A DNA-based molecular device switchable between three distinct mechanical states. Appl. Phys. Lett. 80: 883-885

[43] Yan H, Zhang X, Shen Z, and Seeman NC (2002) A robust DNA mechanical device controlled by hybridization topology. Nature 415: 62-65

[44] Li JJ and Tan W (2002) A single DNA molecular nanomotor. Nano Lett. 2: 315-318

[45] Alberti P and Mergny JL (2003) DNA duplex-quadruples exchange as the basis for a nanomolecular machine. PNAS 100: 1569-1573

[46] Feng L, Park SH, Reif JH, Yan H (2003) A two-state DNA lattice switched by DNA nanoactuator. Angew. Chem. Int. Ed. 42: 4342-4346

Appendix 1: Domain-based Sequence Design of DNA

Throughout the entirety of this thesis, I've emphasized repeatedly that the designs presented should work for most choices of sequences, because we are utilizing the hybridization properties of DNA. While I do believe this to be the case, some sequences clearly work better than others. For substrates and gates that are intended react with naturally-occurring nucleic acid target molecules, there is only limited space for sequence design. But for all synthetic DNA strands and complexes, proper sequence design can help optimize kinetics and minimize unintended side reactions.

A good many DNA sequence design software have been programmed and are in use today. Many of these use only the thermodynamics of potential spurious hybridization to judge the goodness of oligonucleotide sequences. This approach is likely to be suboptimal when designing DNA sequences for dynamic, strand-displacement based reactions and networks. Additionally, there are other considerations that must be taken into account during sequence design, such as the expected concentrations of strands and domains and non-Watson Crick interactions.

Domain sequence design is a complicated problem that a smart student could easily spend his entire graduate career on. My time on this project was limited, so I endeavored in this paper to outline the issues that I believe are important. It was difficult to quantitate the

considerations against one another; in the case of my Domain Design code, I assigned ad hoc values to many parameters. It is worth mentioning, however, that the sequences generated by Domain Design were often "good enough" at the scales I worked with–the sequences I used for my later projects were all designed using Domain Design. More advance sequence design software would likely be needed when hundreds of different oligonucleotides must simultaneously coexist in solution.

This work was submitted to the 16th annual DNA Computing and Molecular Programming Conference as:

Zhang, David Yu. "Towards Domain-based Sequence Design for DNA Strand Displacement Reactions."

DNA strand displacement has been used to construct a variety of components, devices, and circuits. The sequences of involved nucleic acid molecules can greatly influence the kinetics and function of strand displacement reactions. To facilitate consideration of spurious reactions during the design process, one common strategy is to subdivide DNA strands into domains, continuous nucleic acid bases that can be abstracted to act as a unit in hybridization and dissociation. Here, considerations for domain-based sequence design are discussed, and heuristics are presented for the sequence design of domains. Based on these heuristics, a randomized algorithm is implemented for sequence design.

Introduction

DNA strand displacement is a process through which a single-stranded DNA molecule (*strand*) reacts with a multi-stranded DNA *complex* to release another DNA strand (Fig. A1-1AB). Typically, strand displacement is facilitated by *toeholds*, short complementary single-stranded domains that act to colocalize the invading strand with the complex. The thermodynamics of the toehold region can be calculated based on sequence [1–3], and largely determine the kinetics of the strand displacement reaction if the invading strand does not

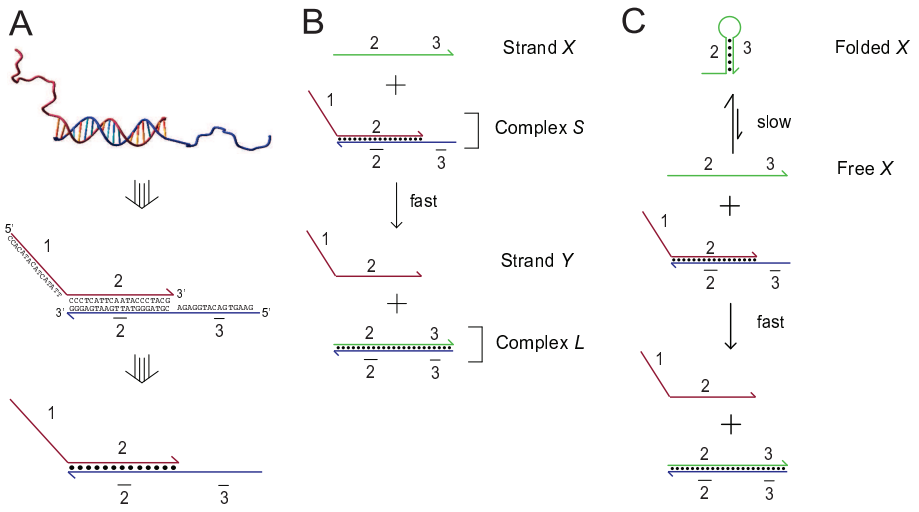


FIG. A1-1: DNA strand displacement **(A)** DNA abstraction. DNA strands and complexes are represented by directional lines, with the hook denoting the 3' end. DNA strands can be functionally abstracted into domains, consecutive bases of DNA that act as a unit in binding and dissociation. Domains are represented by numbers; a barred domain denotes a domain complementary in sequence to the unbarred domain (e.g. domain $\bar{2}$ is complementary to domain 2). Domain 2 is referred to in this paper as a "branch migration domain" and domain $\bar{2}$ is referred to as a "complement domain." Domains 3 and $\bar{3}$ are referred to as "toehold domains." **(B)** DNA strand displacement. **(C)** Secondary structure could hinder the kinetics of strand displacement. Strand X needs to unfold into a free state before it can branch migrate.

possess significant secondary structure [4].

DNA strand displacement has been used to construct a number of dynamic DNA devices, including logic gates and circuits [5–9], catalytic reactions and networks [10–16], and nanoscale motors and walkers [17–20]. These devices' mechanisms are based on Watson-Crick complementarity, and are expected to function for a wide variety of DNA sequences. Nevertheless, the kinetics of these devices' function are slowed if the sequences possess significant secondary structure (Fig. A1-1C) [21, 22]. As strand displacement-based DNA devices become more reliable, many different such devices will be integrated to construct complex reaction networks with more advanced functions. Many different DNA molecules must thus perform their function simultaneously without interfering with each other. Thus, an automated method for DNA sequence design is needed for strand displacement-based DNA devices.

Although excellent algorithmic methods for DNA sequence design have been presented [23–28], these methods generally maximize the probability of DNA strands forming desired structures and complexes at equilibrium, such as in the case for many DNA self as-

sembly applications. For strand displacement-based devices and networks, thermodynamics-based methods of sequence design are not guaranteed to provide satisfactory sequences, because they neglect the consideration of kinetic pathways between DNA strand and complex states. In Fig. A1-2, for example, the strand shown in Fig. A1-2A may proceed through strand displacement slower than the strand in Fig. A1-2B, despite being the former possessing a minimum free energy (mfe) structure with standard free energy (ΔG°) closer to 0.

This paper presents a domain-based approach to sequence design of strand displacement-based reaction, networks, and devices. Domains are consecutive bases that serve as functional units in binding and dissociation (Fig. A1-1A), providing a useful abstraction for the design DNA devices. The sequences of involved DNA strands are obtained by concatenating the sequences of the strand's constituent domains. This domain-based approach to sequence design is simple and generalizes to many different possible DNA complexes and reactions, but is limited in its ability to eliminate undesirable spurious hybridization between DNA strands, particularly at the interface between different domains.

The other extreme, sequence design based on thermodynamic and kinetic analysis of all possible intermediates in strand displacement reactions, would potentially allow a rigorous method for generating optimal sequences for any reaction network. However, such a sequence designer would require significantly more computational resources, and is not available at present time. Additionally, although the thermodynamics of most DNA structural motifs have been carefully characterized over the past 20 years [1–3], two particular ones relevant to the analysis of multi-stranded intermediate complexes remain elusive: the energetics of pseudo-knotted complexes [29–31] and coaxial stacks [32–35]. This incompleteness of DNA thermodynamic data suggest that even these more complex methods may not generate truly "optimal" sequences.

Considerations

In this section, considerations for domain-based sequence design are presented. These considerations are considered generally relevant to sequence design for DNA constructions

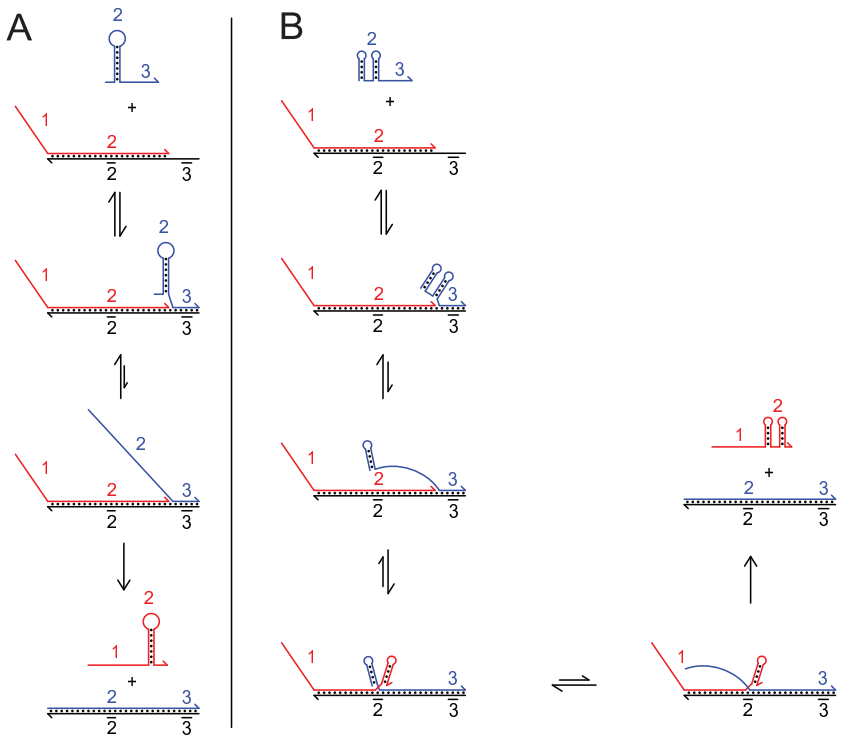


FIG. A1-2: Failure of thermodynamics-based sequence design. The ΔG° (standard free energy) of the folded domain 2 is more negative in **(B)** than in **(A)**, but the sequential nature of branch migration means that the strand displacement reaction involving **(B)** could be faster than that of **(A)**.

involving strand displacement.

Avoidance of long continuous regions of spurious hybridization. Long continuous regions of spurious binding in the branch migration domain should be avoided with priority over several shorter regions of spurious binding, even if the latter results in a more negative minimum free energy structure for the domain or strand. The reason for this is because branch migration is a sequential process: the binding energy of the strongest continuous region of hybridization likely determines the activation energy of the branch migration process.

When multiple different helices are present as in Fig. A1-2B, only the first of them needs to spontaneously open in order for branch migration to initiate. In contrast, when long continuous regions of spurious hybridization exist, branch migration over each of these structured bases is energetically unfavorable, and the kinetics of the overall branch migration process will be slowed exponentially in the energy of the binding region. To take a numerical

example, if the hairpin shown in Fig. A1-2A has energy of −10 kcal/mol, while each of the hairpins in Fig. A1-2B has energy −7 kcal/mol, the kinetics of the strand displacement reaction may be a factor of 100 faster for reaction in Fig. A1-2B at room temperature, despite having an strand mfe of −14 kcal/mol.

Furthermore, because the displaced domain is the same sequence as the branch migration domain, the displaced portion of the domain will start forming the structures that spontaneously opened in branch migration domain. For example, in the bottom-left panel of Fig. A1-2B, the red strand forms the hairpin on its 3' end, analogous to the 3' hairpin of the blue strand. The formation of secondary structure by the displaced strand thus can facilitate branch migration.

There are other possible mechanisms for branch migration through highly structured domains of DNA, such as through 4-way branch migration. However, the author expects that most sequences generated with intent to avoid spurious hybridization will exhibit low enough amounts of spurious hybridization that the dominant pathway for branch migration will be through spontaneous dissociation of spurious hybridization.

Interaction and crosstalk: domain concentration effects. We define a "branch migration domain" to be a domain which competes to binding to a "complement domain" via branch migration. In Fig. A1-1, the 2 domain is a branch migration domain, while the $\bar{2}$ domain is a complement domain. For consistency, branch migration domains are all represented by unbarred numbers.

In strand displacement reactions, there is a necessary excess of branch migration domains over complement domains. For example, there are two copies of domain 2 in Fig. A1-2, but only one copy of domain $\bar{2}$. This excess manifests as a very large difference in the concentrations of single-stranded branch migration domains and complement domains. Almost all of the complement domains will be double-stranded at all times, while the branch migration domains will be single-stranded in significant concentrations.

Spurious hybridization occurs only between two single-stranded domains; this causes potential spurious hybridization involving a complement domain to be significantly less

likely, and that between two complement domains to be nearly non-existent. Consequently, sequence design should primarily seek to minimize spurious binding between different branch migration domains (here refered to as *interactions*, Fig. A1-3B). Spurious hybridization between branch migration domains and complement domains (here refered to as *crosstalk*, Fig. A1-3B) is also undesirable insofar as two different domains may non-specifically displace each other in binding to their respective complements. However, mismatches destabilize the thermodynamics and also significantly impede the kinetics of branch migration [13, 36]. Because most domain sequences designed will differ from each other by at least a few bases, crosstalk is usually not a problem in practice.

Of interactions, self-interactions (wherein a domain significantly hybridizes to an identical copy of itself) can be considered the most problematic for two reasons: First, intra-domain and intra-molecular hybridization are entropically favored because of high local concentration. Second, assuming the thermodynamics of all interactions to be equal, dimerization is likely to be more prevalent than other interactions, because the concentration of a single-stranded domain correlates perfectly with itself (while different domains may not be single-stranded in high concentration at the same time in dynamic circuits).

Thus, sequence design for branch migration domains should avoid domain sequences with self-interactions with highest priority, other interactions with secondary priority, and minimize crosstalk only insofar as it does not hurt the previous two criteria. For sequence design of toehold domains, crosstalk is a larger consideration on the par of interactions.

Minimizing guanine frequency in branch migration domains. Of the four canonical DNA nucleotides, guanine (G) stands out in being the most problematic with regards to sequence design for synthetic biology purposes, both because it is promiscuous (binding to thymine (T) nearly as strongly as adenine (A) [1]) and because it can form guanine quartets [37], a quadruplex structure based on non-Watson Crick binding, usually requiring at least 4 consecutive G's.

For these reasons, it is desirable to minimize the total concentration of guanine nucleotides used in the system. As pointed out previously, branch migration domains are nec-

essarily in higher concentrations than complement domains–this leads to a natural strategy of minimizing the frequency of G's in branch migration domains. This strategy is further desirable because both the branch migration domains and the complement domains are now constructed with only 3 of the 4 nucleotides (C/A/T for the former, G/A/T for the latter), which drastically reduces the chance of self-interactions.

The idea of using only some of the four nucleotide bases in designing DNA sequences was first proposed by Mir [39], who suggested that restricted alphabets may practically avoid nonspecific hybridization, based on experimentally observed hybridization behavior of various sequences [38]. Experimental use of DNA strands with restricted alphabets for DNA nanotechnology purposes is relatively recent [4, 10, 40].

Avoidance of long A/T and long G/C regions. For certain applications, it may be desirable to avoid long uninterrupted stretches of weak (A/T) bases or strong (G/C) bases. Long continuous regions of weak A/T bindings will melt at significantly lower temperatures and breathe significantly more than those with more mixed base distributions. Additionally, the hybridization rate constant of domains with only A/T bases have been reported to be an order of magnitude lower than that of one with a uniform distribution of G/C/A/T bases [4].

On the other hand, long continuous regions of strong G/C bindings are more likely to be spuriously hybridized to other strands and domains, because the strong G/C binding will counteract the destabilizing influences of DNA bulges and mismatches [1]. Additionally, long continuous regions of C/G sequence are more likely to adopt Z-DNA configurations [41]. Finally, branch migration is likely to proceed more slowly in G/C rich regions due to their stronger base stacking thermodynamics, which may in turn necessitate stronger toehold domains for fast strand displacement [4].

Breathing near the end of helices One frequent concern in the design and construction of strand displacement reactions and networks is blunt end strand exchange, strand displacement in the absence of single-stranded toehold domains. The rate constant of blunt end strand exchange has been reported to be on the order of 1 M^{-1} s^{-1} for multiple se-

quences [4, 42, 43]. This rate constant could be significantly higher if the branch migration domain is terminated with A or T.

The mechanism for blunt end strand exchange is postulated as the following: The base pairs at the end of a double-stranded branch migration domain "breathes," temporarily unbinding despite the bound state being more favorable. Any strands possessing the branch migration domain that happen to be in the local vicinity effectively have a few bases of toehold until the ends of the complex re-hybridizes. Terminating branch migration domains with A/T base pairs thus is likely to increase the rate constant of blunt end strand exchange because A/T bases are weaker than G/C ones. Furthermore, measured thermodynamic parameters show that DNA helices are destabilized when closed by an A-T base pair (by 0.15 kcal/mol at room temperature) [1], in addition to the weaker binding thermodynamics.

Consequently, it is desirable for all branch migration and complement domains to start and end with G or C nucleotides. For reasons given previously, it is recommended that branch migration domains start and end with C's.

Interface between domains. One problem specific to domain-based sequence design is the interface between domains. Typically, a stretch of binding requires 3-4 consecutive complementary bases in order to be stable, and such binding will be not be visible to the sequence design software if the bases span multiple domains. Rather than abandoning domain-based sequence design altogether, it should be possible to weight spurious hybridization near the ends of the domains appropriately so that concatenated domains are unlikely to form long spurious hybridization regions at their interface.

Consideration of the potential interface problems through calcuating the interactions and crosstalks for all pairwise concatenations of domains is not recommended for two reasons: first, this causes a runtime slowdown quartic in the number of domains (N^4 pair-wise folds needed for N^2 pair-wise concatenations, where N is the original number of domains), negating the speed advantage of low-level domain-based design software. Second, many of the possible pairwise concatenations will not actually be present in solution, and optimizing the interaction and crosstalk potential of these non-existent domain combinations will

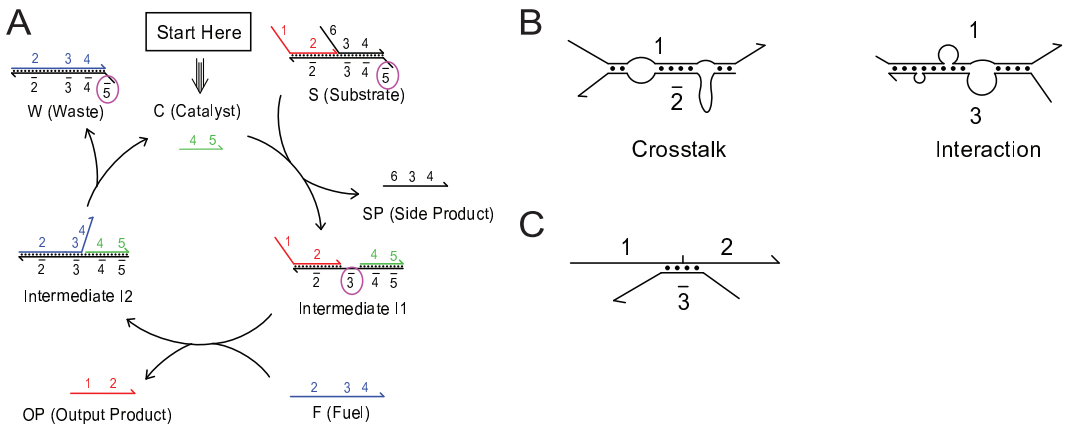


FIG. A1-3: Spurious partial hybridization between unrelated domains. **(A)** Asymmetry of single-stranded prevalence. In the catalytic reaction cycle (adapted from Zhang et al. [10]), the only barred domains that appear in single-stranded forms are $\bar{3}$ and $\bar{5}$, the toeholds. **(B)** Accordingly, we draw a distinction between "crosstalk" and "interaction," with the former denoting spurious partial hybridization between an unbarred domain and a barred domain, while the latter refers to the hybridization between two unbarred domains. Interactions are likely to have greater effects on kinetics than crosstalk. **(C)** One shortcoming of domain-based sequence design is the possibility of crosstalk and/or interactions at the interface between two domains.

actually worsen the crosstalk and interaction problems of the strands that do exist.

Implementation

The Domain Design (DD) software presented here employs heuristics to quantify the considerations discussed in the previous section; these heuristics are used to generate an overall "score" for each domain, the worst (maximum) of which is the global score for a set of domains. Simply put, the score is a metric that roughly correlates with the likelihood that the kinetics of strand displacement reactions deviate from predictable models [4]. The overall score for a set of domains is simply the maximum (worst) of the domain scores. DD's algorithm performs sequence design by starting with a random set of sequences, and continually attempting to mutate these sequences to achieve improved (lower) scores.

DD assumes that the domains designed are branch migration domains or the unbarred toehold domains, and evaluates crosstalk potential by automatically generating the complements to all designed domains. One good design strategy may be to design all the toehold domains first, and subsequently design the branch migration domains (with the toehold domain sequences locked–see User Interactivity: Base locking).

The algorithm uses a number of scoring parameters, some of which are hard-coded and can be turned on or off at the user's discretion. The values of the hard-coded parameters are by no means guaranteed to be optimal or even close to optimal–these represent only the author's best guess at good parameter values for designing 10 or fewer different domains, each of length between 5 and 30 nucleotides. The pseudocode for the Domain Design software is given in Fig. A1-4.

Score calculation. The score of a domain is computed as the sum of the domain intrinsic score and the worst of its crosstalks and interactions.

The domain intrinsic score accounts for score penalties based only on the sequence of the domain, rather than potential spurious hybridization. Domains with four consecutive G's or C's receive a +50 to score, so as to strongly discourage the formation of G quartets. Domains with six consecutive G/C nucleotides or six consecutive A/T nucleotides receive a +20 to score, if the user elects to turn on the option for avoiding long regions of G/C binding and long regions of A/T binding. Finally, the user-defined "importance" of the domain is also added to the score (see User Interactivity: Domain importance).

The crosstalk and interaction scores are evaluated similarly, by evaluating every potential way that the two domains can spurious hybridize and identifying the way that yields the highest score. The score of each way of binding is calculated as thus: each G-C base pair contributes +2 score and each A-T pair contributes +1 score. Base pairing of x consecutive nucleotides without intervening bulges or mismatches gain a further score of $+2^{x-4}$ for $x \geq 5$, thus strongly discouraging the formation of long unbroken stretches for spurious hybridization. Mismatch and bulge structures between stretches of hybridization each contribute -3 score, with an additional -0.5 score for every base in the bulge or mismatch in excess of 1. The scores of hybridization segments starting at the first base of a domain or ending with the last base of a domain are increased by +3, to reduce the potential for interface problems when concatenating domains.

The difference in DD's treatment of crosstalk, interactions, and self-interactions is that each's score is modified linearly. Self-interactions receive a +5 to score, interactions default

to no adjustment, and crosstalk scores are divided by 2 and then a -10 to score is applied. Thus, DD places higher priority on self-interactions and lower priority on crosstalk.

In practice, DD will optimize the sequences of a set of 10 domains, each 20 nt long, in about a minute to the point where the overall score of the system is between +10 and +15.

Algorithm. DD calculates the interaction and crosstalk score of a pair of domains by using a dynamic programming algorithm similar to those used mFold, DINAMelt, and PairFold [3, 44, 45]. See source code for details.

In every run of the main loop, DD attempts to improve the overall score by mutating one of the domains. If the overall score was improved through the mutation, DD keeps the mutation. If the overall score was unchanged through the mutation, DD keeps the mutation with 0.2 probability. If the overall score was worsened through the mutation, DD discards the mutations. At the user's option, DD will either randomly select one of the domains for attempted mutation with uniform probability, or will target the domain currently with the worst score with probability $\frac{1}{3}+\frac{2}{3N}$, where N is the number of domains designed. Targeting the domain with worst score for mutations is expected to improve the speed of the software.

The number of bases attempted to be mutated is roughly exponentially distributed, with $\frac{1}{2}$ probability of mutating 1 base, $\frac{1}{4}$ of mutating 2 bases, etc., up to 10 bases or M (the number of bases in the domain), whichever is smaller. The simultaneous mutation of multiple bases is thought to prevent the domain sequences from entering local score optima.

Each base attempted to be mutated is replaced by a random base drawn from the bases allowed to occur within the domain (see User Interactivity: Base Composition). This means that there is some probability that an attempted mutation does not actually change the base, because the new base happens to be the same as the old. At the user's option, the mutation process can be biased against the incorporation of G, so that only 4% of attempted mutations result in a G (assuming G is an allowed base in the first place).

Time and space complexity. The time complexity of the algorithm can be estimated easily. Define the problem to be the design of N different domains, each of length M. After every mutation attempt, the software must update the interaction and crosstalk scores

```
FOR i = 1:N
  FOR j = 1:N
    Score(i,j) = Domain_intrinsic_score(i) + Max(Crosstalk(i,j), Interaction(i,j))
  END FOR
  Domain_Score(i) = Max_over_j(Score(i,j))
END FOR
Global_score = Max_over_i(Domain_Score(i))

WHILE (TRUE)
  k = Random(1,N)
  New_domain(k) = Mutate(Domain(k))

  FOR i = 1:M
    New_Score(i, k) = Domain_intrinsic_score(i) + Max(Crosstalk(i,k), Interaction(i,k))
    New_Score(k, i) = New_Score(i,k)
    IF (i != k)
      Domain_Score(i) = Max_over_j(Score(i,j))
    END IF
  END FOR
  Domain_Score(k) = Max_over_j(Score(k, j)
  New_global_score = Max_over_i(Domain_Score(i))

  IF (New_global_score < Global_Score)
    Domain(k) = New_domain(k)
    FOR i = 1:M
      Score(i,k) = New_score(i,k)
      Score(k,i) = New_score(k,i)
    END FOR
    Global_Score = New_global_score
  END IF

  IF (Keyboard_input())
    Prompt_user_menu()
  END IF
END WHILE
```

FIG. A1-4: Pseudo-code for the Domain Design software.

of all domains with the mutated domain, which involves checking O(N) interactions and crosstalks. Calculating an interaction or crosstalk score using dynamic programming requires O(M^2) time. Thus, the algorithm has time complexity O(NM^2) per mutation attempt.

The number of total mutation attempts needed to in order to grant each domain a fixed expected number of mutation attempts will scale linearly with N. Thus, DD requires O(N^2M^2) time to reduce the overall score to decent levels.

In addition to the O(NM) space needed to store the sequences of the domains, there are two major contributions to the space complexity of the algorithm. First, during the evaluation of a crosstalk or interaction score, O(M^2) space is needed for the dynamic programming algorithm. Second, there are O(N^2) pairwise crosstalk and interaction scores between the N domains; these scores need to be stored in order to allow fast score updating

(not storing the scores worsens the time complexity by a factor of N). Thus, the overall space complexity of the algorithm is $\mathrm{O}(N^2 + NM + M^2) = \mathrm{O}(N^2 + M^2)$.

User Interactivity

The Domain Design software (DD) presented here was written with the intention of being a tool for helping the informed DNA engineer design sequences, rather than being an arbitrator of sequences. Accordingly, Domain Design strives to maximize the ease with which the user can modify and adjust the sequence design process: The user may pause the sequence optimization process at any time to tweak sequences as well as a number of design parameters, including the ones listed below. User interactive sequence design is considered advantageous because the user may be able to identify sequence problems that the software is not able to detect.

Base locking. DD acknowledges that the user may possess certain constraints or preferences in the design of sequences. For example, the user may require the promoter sequence for the T7 RNA polymerase [46, 47] or a deoxyribozyme sequence [48–50] to be present at a given location. To this end, the user can manually change the sequences of any domain, and can furthermore lock part or all of a domain sequence from being mutated by the software. Locked bases are visually displayed in red (Fig. A1-5A).

Domain importance. Different domains serve different purposes in a reaction network, and the frequency of a domain being single-stranded also varies among domains. For example, toehold domains are required to colocalize the reactants of the strand displacement reaction, and interactions/crosstalk involving toeholds may slow the kinetics of strand displacement much more than similar interactions/crosstalk between branch migration domains. Consequently, sequence-based interactions and crosstalk in the different domains need to be weighed differently. A rigorous and justified method for weighing the domains must take into account the the context of the overall reaction network, and is beyond the scope of a low-level domain-based approach to sequence design.

Instead, DD implements a user-defined "importance" parameter, which acts as an ad-

ditive term to the score of the domain (Fig. A1-5B). In DD, the overall score of the set of domain sequences is determined by the domain with the worst (highest) score, and in a typical run the scores of all of the domains will be similar. The additive importance term to score means that high importance domains must have lower values for the other score terms in order to have a similar score to domains with low importance.

Base composition. DD allows the user to specify the base compositions of each of the domains. For reasons outlined previously, it may be desirable to design certain domains using 3-letter C/A/T alphabets. Furthermore, the user also may wish to restrict certain domains to only A/T bases or G/C bases, so that the domain is weak- or strong-binding. This feature also facilitates the sequence design of non-standard DNA structures, such as Z-DNA [41] (C/G), or Hoogsteen triplex bindings (C/T) [51].

Designing sequences to function with an existing system. Finite research funding usually implies that DNA nanotechnologists will tend to use and reuse the same DNA oligonucleotide strands for many different purposes. It is not only economical but also practical, because using previously tested strands eliminates the chance for unexpected sequence-specific problems. Given this tendency, it is therefore important for sequence design software to be able to design optimal sequences given the constraints of the sequences of the existing strands.

DD allows the user to load sequences from files before attempting to design new sequences. At the user's option, the domains loaded in this manner can be all be locked for the new design process. The user may also save the sequences generated by DD at any time, and the save file will include information such as the importance, composition, and lock status of all bases and domains.

Discussion

This paper discusses the author's considerations for sequence design of DNA strand displacement-based reactions and cascades, and makes a tentative relative weighing of the considerations in the form of various score penalties. A randomized algorithm was presented

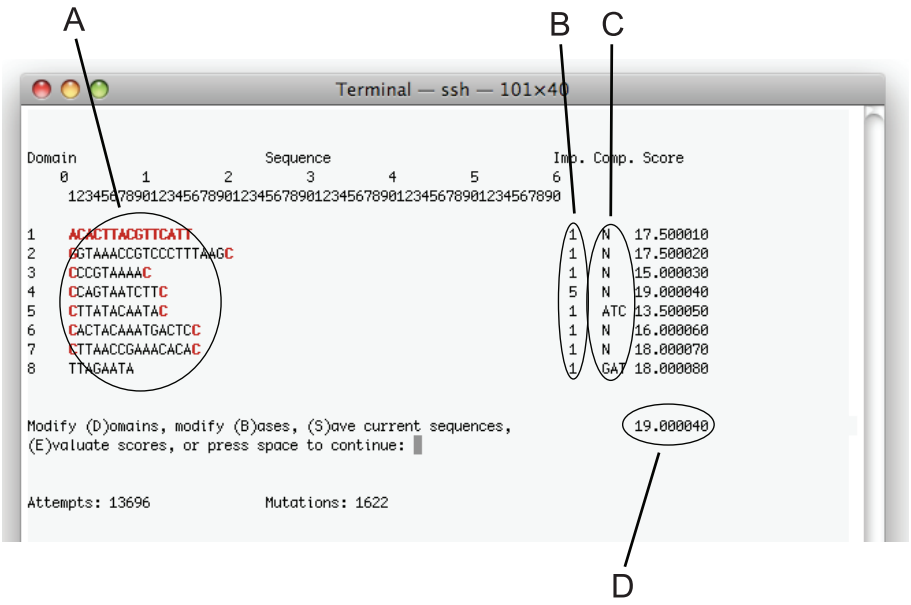


FIG. A1-5: Sample run of Domain Design (DD) software. The sequences of the domains are shown in **(A)**, with the bases in red denoting bases that are "locked" and prevented from mutation. The user can lock or unlock bases and domains through the course of the design process, as well as manually change the sequences of the domains. The numbers in the **(B)** column list the "importance" of each domain, which is implemented as an additive adjustment to the score of the domain. **(C)** shows the nucleotide compositions of each of the domains, with "N" denoting no constraints (all four nucleotides may be used). The overall score of the set of domains is shown in **(D)**. The tiny decimal portion in the overall score shows the number of the domain currently possessing the worst score, to facilitate user intervention. In the figure, Domain 4 currently has the worst score.

that attempts to minimize the overall score of the domains to be designed by mutating a few bases at a time, and accepting mutations that lead to improved (lower) scores. The Domain Design (DD) software runs the randomized algorithm for sequence design while allowing the user to dynamically intervene.

At its heart, DD is a low-level domain-based approach to sequence design, and does not consider the ways in which the domains are concatenated to form DNA strands, nor the overall architecture of the reaction network using those strands. As a result, the sequences generated by DD will in general not be as good as those generated by software intended specifically for particular applications. The comparative advantage of DD is its speed, simplicity, and generalizability–by ignoring the details of particular strands and systems, it seeks to generate sequences that are "good enough" for a wide variety of DNA nanotechnology applications based on strand displacement. Empirical evidence of the success of DD lie in experimental works published by the author, who used DD to design

the sequences for a variety of strand displacement-based reactions and networks exhibiting catalysis [4, 8, 12, 13].

Currently, most experimental work on dynamic DNA nanotechnology have been very limited in scale, with numbers of different DNA strands concurrently in solution being less than 100. At these scales, it is relatively easy to design sequences that do not significantly interact or crosstalk with one another. As a result, many different sequence design software (both published and unpublished) will likely generate sequences that work decently well for their intended applications, and it is not easy to critically evaluate their relative performances. As researchers develop and experimentally test larger reaction networks [40, 52, 53], the demand for large numbers of DNA sequences that must exist stably simultaneously in solution will drive the development of ever better and easy-to-use DNA sequence design software.

Source code for Domain Design can be downloaded at:
http://www.dna.caltech.edu/∼dzhang/softsource/DD.c

Acknowledgements: This research is supported in part by NSF grants 0832824 and 0728703. DYZ is supported by the Fannie and John Hertz Foundation.

The author thanks Erik Winfree and David Soloveichik for testing the DD code. The author thanks Erik Winfree for suggestions regarding the manuscript.

[1] J. SantaLucia, D. Hicks, *Annu. Rev. Biophys. Biomol. Struct.* **33,** 415 (2004).

[2] R. M. Dirks, J. S. Bois, J. M. Schaeffer, E. Winfree, N. A. Pierce, *SIAM Rev.* **49,** 65 (2007).

[3] Zuker, M. *Nucleic Acids Res.* **31,** 3406 (2003).

[4] D. Y. Zhang, E. Winfree, *J. Am. Chem. Soc.* **131,** 17303 (2009).

[5] G. Seelig, D. Soloveichik, D. Y. Zhang, E. Winfree, *Science* **314**, 1585 (2006).

[6] M. Hagiya, S. Yaegashi, K. Takahashi, *Nanotechnology: Science and Computation*, 293-308 (2006).

[7] B. M. Frezza, S. L. Cockroft, M. R. Ghadiri, *J. Am. Chem. Soc*, **129**, 14875 (2007)

[8] D. Y. Zhang, "Cooperative DNA strand displacement for DNA quantitation, detection, and logic." (submitted, 2010).

[9] Z. Xie, S. J. Liu, L. Bleris, Y. Benenson, *Nuc. Acids Res.* (2010, DOI: 10.1093/nar/gkq117).

[10] D. Y. Zhang, A. J. Turberfield, B. Yurke, E. Winfree, *Science* **318**, 1121 (2007).

[11] A. J. Turberfield, J. C. Mitchell, B. Yurke, A. P. Mills, M. I. Blakey, F. C. Simmel, *Phys. Rev. Lett.* **90**, 118102 (2003).

[12] D. Y. Zhang, E. Winfree, *J. Am. Chem. Soc.* **130,** 13921 (2008).

[13] D. Y. Zhang, E. Winfree, *Nuc. Acid Res.* (2010, pre-published online DOI:10.1093/nar/gkq088).

[14] G. Seelig, B. Yurke, E. Winfree, *J. Am. Chem. Soc.* **128,** 12211 (2006).

[15] J. S. Bois, S. Venkataraman, H. M. T. Choi, A. J. Spakowitz, Z. G. Wang, N. A. Pierce, *Nuc. Acid Res.* **33,** 4090 (2005).

[16] S. J. Green, D. Lubrich, A. J. Turberfield, *Biophysical Journal* **91,** 2966 (2006).

[17] B. Yurke, A. J. Turberfield, A. P. Mills, F. C. Simmel, J. L. Neumann, *Nature* **406**, 605 (2000).

[18] R. M. Dirks, N. A. Pierce, *Proc. Nat. Acad. Sci.* **101**, 15275 (2004).

[19] P. Yin, H. M. T. Choi, C. R. Calvert, N. A. Pierce, *Nature* **451**, 318 (2008).

[20] T. Omabegho, R. Sha, N. C. Seeman, *Science* **324,** 67 (2009).

[21] Y. Gao, L. K. Wolf, R. M. Georgiadis, *Nuc. Acids Res.* **34,** 3370 (2006).

[22] W. Sun, C. Mao, F. Liu, N. C. Seeman, *J. Mol. Biol.* **282,** 59 (1998).

[23] R. M. Dirks, M. Lin, E. Winfree, N. A. Pierce, *Nucleic Acids Res.* **32,** 1392 (2004).

[24] J. Seifferf, A. Huhle, *J. Biomol. Struct. Dyn.* **25,** 453 (2008).

[25] F. Tanaka, A. Kameda, M. Yamamoto, A. Ohuchi, *Nuc. Acids Res.* **33,** 903 (2005).

[26] D. Tulpan, M. Andronescu, S. B. Chang, M. R. Shortreed, A. Condon, H. H. Hoos, L. M. Smith, *Nuc. Acids Res.* **33,** 4951 (2005).

[27] J. Sager, D. Stefanovic, *Lecture Notes of Computer Science* **3892,** 275-289 (2006).

[28] N. C. Seeman, *J. Biomol. Struct. Dyn.* **8,** 573-581 (1990).

[29] S. Cao, S. Chen, *Nuc. Acids Res.* **34,** 2634 (2006).

[30] A. Xayaphoummine, T. Bucher, H. Isambert, *Nuc. Acids Res.* **33,** W605 (2005).

[31] R. M. Dirks, N. A. Pierce, *J. Comput. Chem.* **25,** 1295 (2004).

[32] E. Protozanova, P. Yakovchuk, M. D. Frank-Kamenetskii, *J. Mol. Biol.* **342**, 775 (2004).

[33] D. V. Pyshnyi, E. M. Ivanova, *Russian Chemical Bulletin* **51**, 1145 (2002).

[34] V. A. Vasiliskov, D. V. Prokopenko, A. D. Mirzabekov, *Nuc. Acid Res.* **29**, 2303 (2001).

[35] D. V. Pyshnyi, E. M. Ivanova, *Nucleosides, Nucleotides, and Nucleic Acids* **23**, 1057 (2004).

[36] I. G. Panyutin, P. Hsieh, *J. Mol. Biol.* **230,** 413 (1993).

[37] D. Sen, W. Gilbert, *Methods in enzymology* **211,** 191 (1992).

[38] E. M. Southern, S. C. Casegreen, J. K. Elder, M. Johnson, K. U. Mir, L. Wang, J. C. Williams, *Nuc. Acids Res.* **22,** 1368 (1994).

[39] Mir, K. U. A restricted genetic alphabet for DNA computing. *DNA based computers II. DIMACS* **44,** 243-246 (1998).

[40] L. Qian, E. Winfree, *Lecture Notes of Computer Science* **5347,** 70 (2009).

[41] C. Mao, W. Sun, Z. Shen, N. C. Seeman, *Nature* **397,** 144 (1999).

[42] L. P. Reynaldo, A. V. Vologodskii, B. P. Neri, V. I. Lyamichev, *J. Mol. Bio.* **297**, 511 (2000).

[43] B. Yurke, A. P. Mills, *Genet. Prog. Evol. Mach.* **4**, 111 (2003).

[44] M. Zuker, D. H. Mathews, D. H. Turner, "Algorithms and Thermodynamics for RNA Secondary Structure Prediction: A Practical Guide In RNA Biochemistry and Biotechnology," in J. Barciszewski and B.F.C. Clark, eds., NATO ASI Series, Kluwer Academic Publishers, (1999).

[45] R. A. Dimitrov, M. Zuker, *Biophys. J.* **87,** 215 (2004).

[46] J. Kim, K. S. White, E. Winfree, *Mol. Syst. Biol.* **2,** 68 (2006).

[47] W. U. Dittmer, F. C. Simmel *Nano Lett.* **4,** 689 (2004).

[48] Stojanovic, M. N., Semova, S., Kolpashchikov, D., Macdonald, J., Morgan, C. & Stefanovic, D. *J. Am. Chem. Soc.* **127,** 6914-6915 (2005).

[49] R. Pei, S. K. Taylor, D. Stefanovic, S. Rudchenko, T. E. Mitchell, M. N. Stojanovic, *J. Am. Chem. Soc.* **128,** 12693 (2006).

[50] K. Lund, A. Manzo, N. Dabby, N. Michelotti, A. Johnson-Buck, J. Nangreave, S. Taylor, R. Pei, M. N. Stojanovic, N. Walter, E. Winfree, H. Yan, *Nature* (in press, 2010).

[51] M. D. Frank-Kamenetskii, S. M. Mirkin, *Annu. Rev. Biochem.* **64,** 65 (1995).

[52] D. Soloveichik, G. Seelig, E. Winfree, *Proc. Nat. Acad. Sci.* (2010, pre-published online DOI:10.1073/pnas.0909380107).

[53] A. Phillips, L. Cardelli, *Journal of the Royal Society Interface* **6,** S419 (2009).