High-Fidelity DNA Hybridization using Programmable Molecular DNA Devices

Nikhil Gopalkrishnan **Harish Chandran** **John Reif**

Abstract

The hybridization of complementary nucleic acid strands is the most basic of all reactions involving nucleic acids, but has a major limitation: the specificity of hybridization reactions depends critically on the lengths of the complementary pairs of strands and can drop (in the presence of other competing strands whose sequences are close to that of a given target strand) to very low values if the strands have sufficiently long length. This reduction in specificity of hybridization reactions occurs especially in the presence of *noise* in the form of other competing strands that have sequence segments identical to the target. This limitation in specificity depending on strand length significantly limits the scale and accuracy of biotechnology and nanotechnology applications which depend on hybridization reactions. Our paper develops techniques for ensuring specific high-fidelity DNA hybridization reactions for target strands of arbitrary length. Given an *in vitro* solution which contains various DNA strands with differing sequences, among them a particular known target DNA sequence s of relatively long length (say at least 60 to hundreds of bases), our goal is to bind to each subsequence segment of s with high specificity and exact complementarity for a significant fraction of strands s in the solution. To do this, we develop a protocol that relies only on hybridization reactions between relatively short length (of at most 15 bases) DNA sequence segments. Our basic approach is to design DNA devices that essentially scan over strands in solution, subsegment by subsegment, and determine if one is indeed an instance of the given target strand s. This scanning is achieved by carefully designed relatively short sequences that effectively perform, in sequential order, successive verifications of subsequence identities on the target sequence, which if successfully completed indicate that the complete target sequence has been verified and completely hybridized. Our protocol is executed autonomously, without external mediation. Our high-fidelity DNA hybridization protocol is driven by a series of conversions of single stranded DNA into duplex DNA that help overcome kinetic energy traps, similar to DNA walkers. In addition to the design of our high-fidelity DNA hybridization protocol, we also discuss the kinetics of our hybridization reactions, reducing the overall kinetics to a series of well-understood strand-displacement reactions. Further, we describe a detailed design of an ongoing small-scale experimental demonstration of our protocols for high-fidelity hybridization. We also discuss potential applications for our protocols in molecular detection and DNA computing.

1 Introduction

1.1 Motivation

The hybridization of complementary nucleic acid strands is the most basic of all reactions involving nucleic acids and a major component of most protocols involving nucleic acids. Indeed, hybridization reactions are the basis for much of biotechnology involving nucleic acids. For example, they are essential to many DNA enzymatic reactions such as restriction cuts, to PCR reactions used for amplification and for the operation of DNA hybridization arrays. Hybridization reactions are also the basis of DNA nanotechnology, which use hybridization of strands to form DNA tile nanostructures, as well as to bind DNA tile nanostructures together to form DNA lattices and to form DNA origami via hybridization between long scaffold strands and short staple strands.

However, the hybridization reaction has certain key limitations. The primary limitation is that the specificity of hybridization reactions (the likelihood that a given strand only hybridizes with its exact complementary strand) depends critically on the lengths of the complementary pairs of

strands. While hybridization reactions in the appropriate solution conditions and temperature can have high-fidelity for moderate strand lengths (from 5 to 15 bases), the specificity of hybridization reactions can drop to very low values if the strands have sufficiently long length (say 25 or more bases). This reduction in specificity of hybridization reactions occurs especially in the presence of *noise* in the form of other competing strands that have sequence segments identical to the target. This limitation in the specificity of hybridization reactions depending on strand length significantly limits the scale and accuracy of biotechnology and nanotechnology applications which depend on hybridization reactions. For example, it limits the length of PCR primers, the length and thus the number of distinct strands in DNA hybridization arrays and the complexity of certain DNA nanostructures such as DNA origami.

1.2 Problem Statement: High-Fidelity DNA Hybridization

Let ssDNA denote a single stranded DNA and dsDNA denote a duplex DNA. Consider a solution containing distinct DNA sequences, with one of these sequences designated as a *target* s. We assume the particular known target DNA strand s is of relatively long length (say at least 60 to hundreds of bases). A target ssDNA in solution is said to be *completely hybridized* if all bases of the strand are (Watson-Crick) hybridized to corresponding complementary bases on other ssDNA, thus leaving no single stranded region on it. Note that multiple ssDNA may contribute complementary bases and thus cooperatively completely hybridize the target. The problem of *Exact High-Fidelity DNA Hybridization* is to completely hybridize each instance of s in solution while no instances of any other strand is completely hybridized. The problem of Exact High-Fidelity DNA Hybridization appears too stringent to be achievable in practice, since it does not allow for a small probability of failure or incomplete hybridizations nor does it allow for minor base mismatches. Hence we instead will take as our goal an approximate version of the High-Fidelity DNA Hybridization defined as follows.

The *Levenshtein distance* is a metric for measuring the edit difference between two sequences. In this paper the allowable edit operations on DNA sequences are insertion, deletion or substitution of a single base. A ssDNA in solution is *b-hybridized* if all but b bases of the strand are (Watson-Crick) hybridized to corresponding complementary bases on other ssDNA. Note that multiple ssDNA may contribute complementary bases and thus cooperatively b-hybridize the target. Given a fixed *failure probability* p, and *base mismatch error* b (where $0 < p < 1$ and b is a positive constant integer), the problem of (p, b)-*High-Fidelity DNA Hybridization* is to b-hybridize with probability p each instance of the target s in solution while no other strand is b-hybridized with probability greater than $1 - p$ (Note: Typically, in our protocols, $1 - p$ might be in the order of a few percentages and b might be a very small constant). Even this approximate version of High-Fidelity DNA Hybridization is quite challenging, as discussed in the first subsection.

1.3 Our Results: Protocols for High-Fidelity DNA Hybridization using DNA Devices

In this paper, we describe two protocols to achieve high-fidelity hybridization to an arbitrary given target DNA sequence. Our high-fidelity DNA hybridization protocols have the following favorable properties:

- Our protocols use only hybridization reactions of relatively short length (of approximately at most 15 bases), which are inherently highly specific.
- Our protocols are executed autonomously, without external mediation.

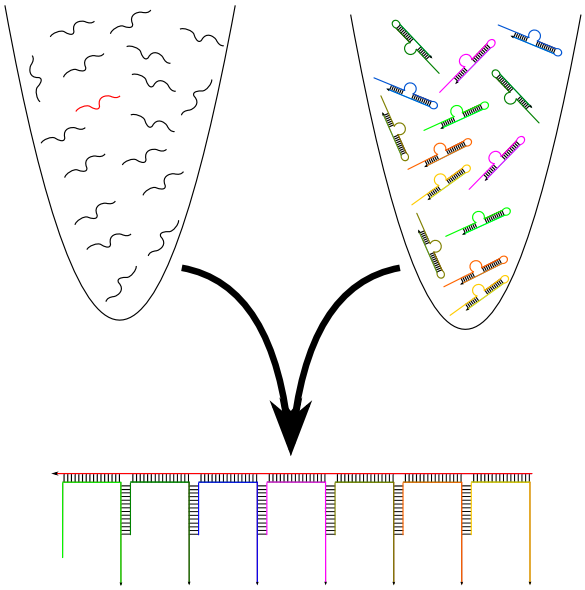

Figure 1: High-fidelity hybridization problem

Our basic approach is to design DNA devices that essentially scan over strands in solution, subsegment by subsegment, and determine if one is indeed an instance of the given target strand s. This scanning is achieved by carefully designed relatively short sequences (called *checker sequences*) that hybridize to distinct contiguous subsequences on the target sequence (see fig. 1). These checker strands perform successive *subsequence verification*. They are designed such that if the appropriate subsequence on one of them doesn't hybridize sufficiently to a specific subsegment of potential target strand, the subsequent *checker* strands do not hybridize to this potential target. Completion of the series of hybridizations indicates that the complete strand has been verified and hybridized.

To ensure protocols are executed autonomously, without external mediation, our high-fidelity DNA hybridization protocols are driven by a series of conversions of single stranded DNA into duplex DNA that help overcome kinetic energy traps, similar to DNA walkers (see Sherman and Seeman (2004); Yin et al. (2004); Tian et al. (2005)). In addition to the design of our high-fidelity DNA hybridization protocol, we also discuss the kinetics of our hybridization reactions, reducing the overall kinetics to a series of well-understood strand-displacement reactions. Further, we describe a detailed design of an ongoing small-scale experimental demonstration of our protocols for high-fidelity hybridization. We also discuss potential applications for our protocols in molecular detection and DNA computing.

1.4 Notation

We facilitate a general symbolic design and description of our protocols and component DNA nanostructures by using the following notational conventions in text and figures:

- All DNA sequences will be represented by Latin letters and these letters may have subscripts, for example c_i.

- Subsequences are also denoted by letters and can also have subscripts, for eg. $a_i b_i = c_i$ where c_i is the concatenation of the subsequences a_i and b_i.

- The sequences are always written from $5'$ to $3'$. Arrows on DNA strands in the figures indicate $3'$ ends.
- Sequences denoted with the same letter that differ only in the subscript will be concatenations of subsequences that only differ in the subscript. For eg. $c_i = a_i d_{i-1}$ implies that the sequence c_{i+1} is a concatenation of a_{i+1} and d_i.
- A *bar* over a sequence indicates reverse complement of a sequence. For eg. $\bar{c}_i$ is the reverse complement of c_i and $\bar{b}_i \bar{a}_i$ is the reverse complement of $a_i b_i$.
- To decrease cluttering of figures, we only indicate the subsequences of one of the component strands of dsDNA.

2 First Protocol for high-fidelity hybridization

Let $s = r_n t_n r_{n-1} t_{n-1} \dots r_i t_i \dots r_1 t_1$ be a long DNA target strand, where r_i and t_i for $i = 1, \dots, n$ are DNA subsequences. We wish to bind short checker sequences $c_i : i = 1, \dots, n$ to s. Figure 2 indicates two consecutive checker sequences c_i in red and c_{i+1} in green. The expected secondary structure of these strands is also indicated in the figure. As an inductive hypothesis, assume that c_i is bound to the blue template strand s as shown in Figure 2. We claim that the appropriate complementary portion of strand c_{i+1}, $\bar{t}_{i+1} \bar{r}_{i+1}$ binds to $r_{i+1} t_{i+1}$ on s and the induction is advanced by one step.

First, u_{i+1} on the green strand binds to its complementary region $\bar{u}_{i+1}$ which is a part of the red strand (Fig. 2). Through this toehold, a strand displacement reaction occurs, breaking the bond between v_{i+1} and $\bar{v}_{i+1}$. v_{i+1} is now attached to its complementary region $\bar{v}_{i+1}$ on the red strand. This opens the hairpin $\bar{y}_i \bar{t}_{i+1}$ allowing $\bar{y}_i$ to attach to y_i on the red strand and also $\bar{t}_{i+1}$ to attach to t_{i+1} on the blue strand (Fig. 3). Now, another strand displacement reaction occurs via the toehold t_{i+1} on the blue strand which breaks the bonds between r_{i+1} and $\bar{r}_{i+1}$ allowing $\bar{r}_{i+1}$ on the green strand to bind to r_{i+1} on the blue strand (Fig. 4). This opens the hairpin $y_{i+1} \bar{v}_{i+2} \bar{u}_{i+2}$. Thus, the green strand is in the same conformation as the blue was at the beginning of the induction step. The sequence $\bar{v}_{i+2} \bar{u}_{i+2}$ on the green strand can now open up the first hairpin on strand c_{i+2}, thus activating it and the process continues till all c_i are bound to the target s. If the potential target s does not have the appropriate sequence, some checker strand will not bind and hence the sequence of attachments will be halted.

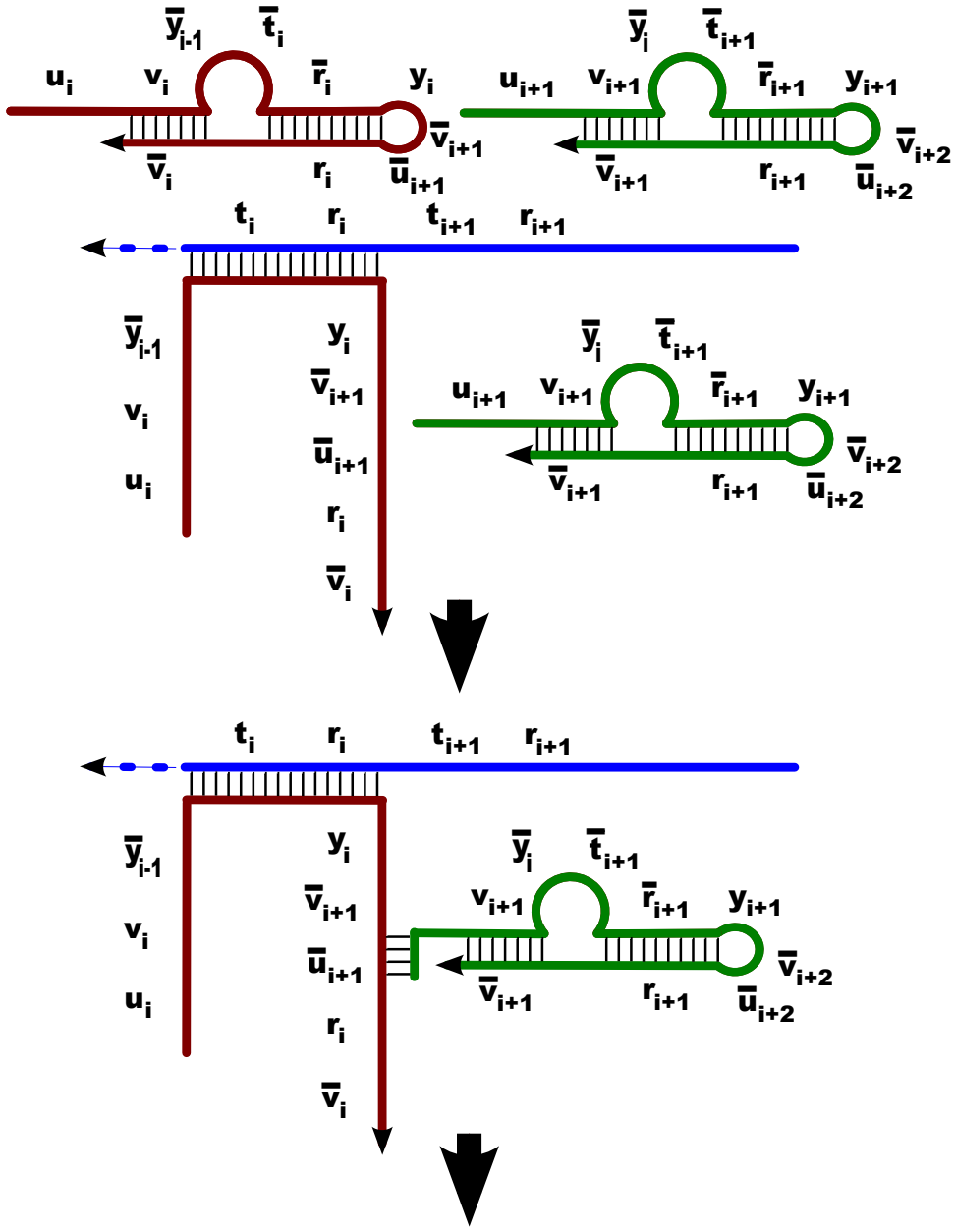

Figure 2: First protocol - I

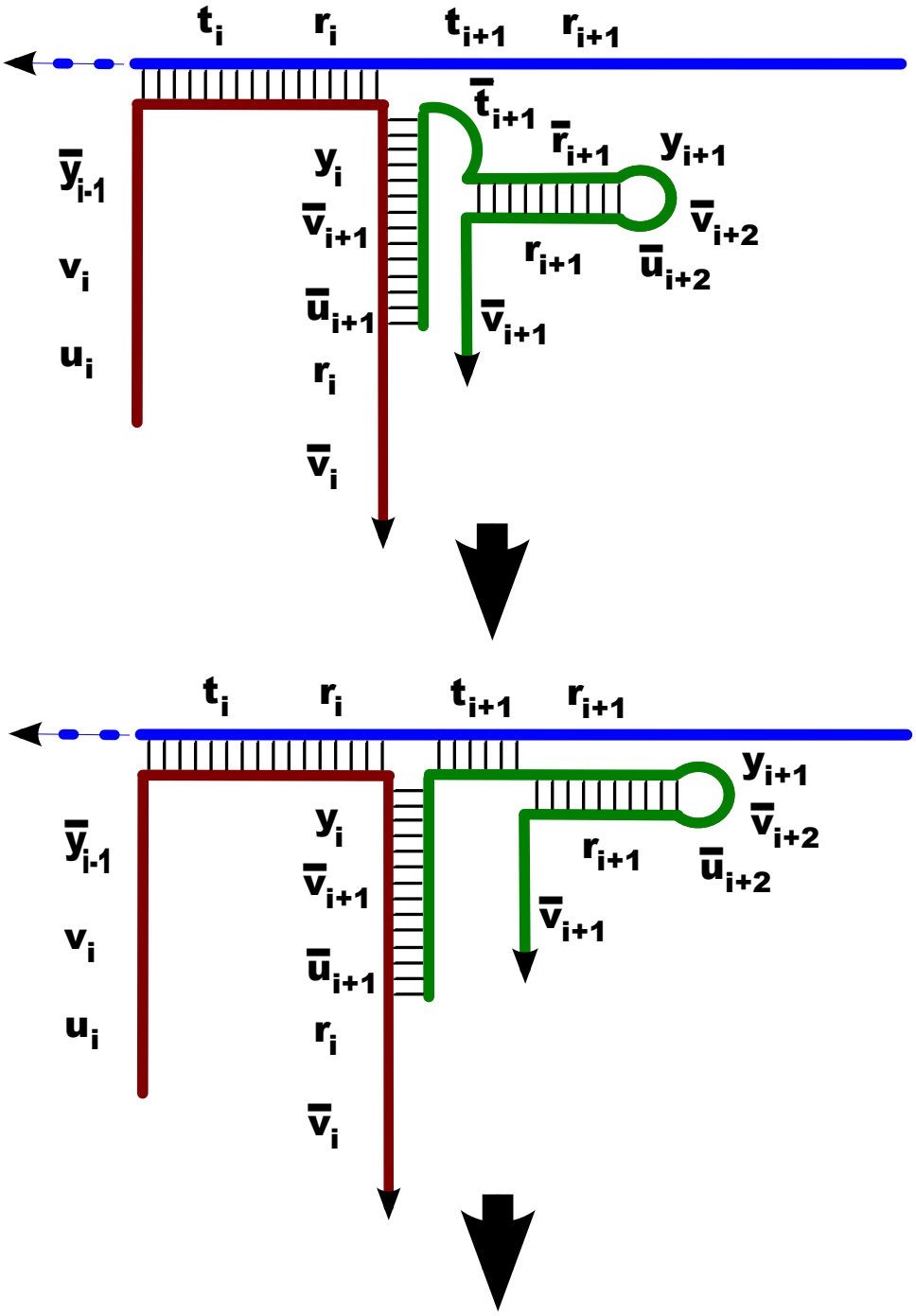

Figure 3: First protocol - II

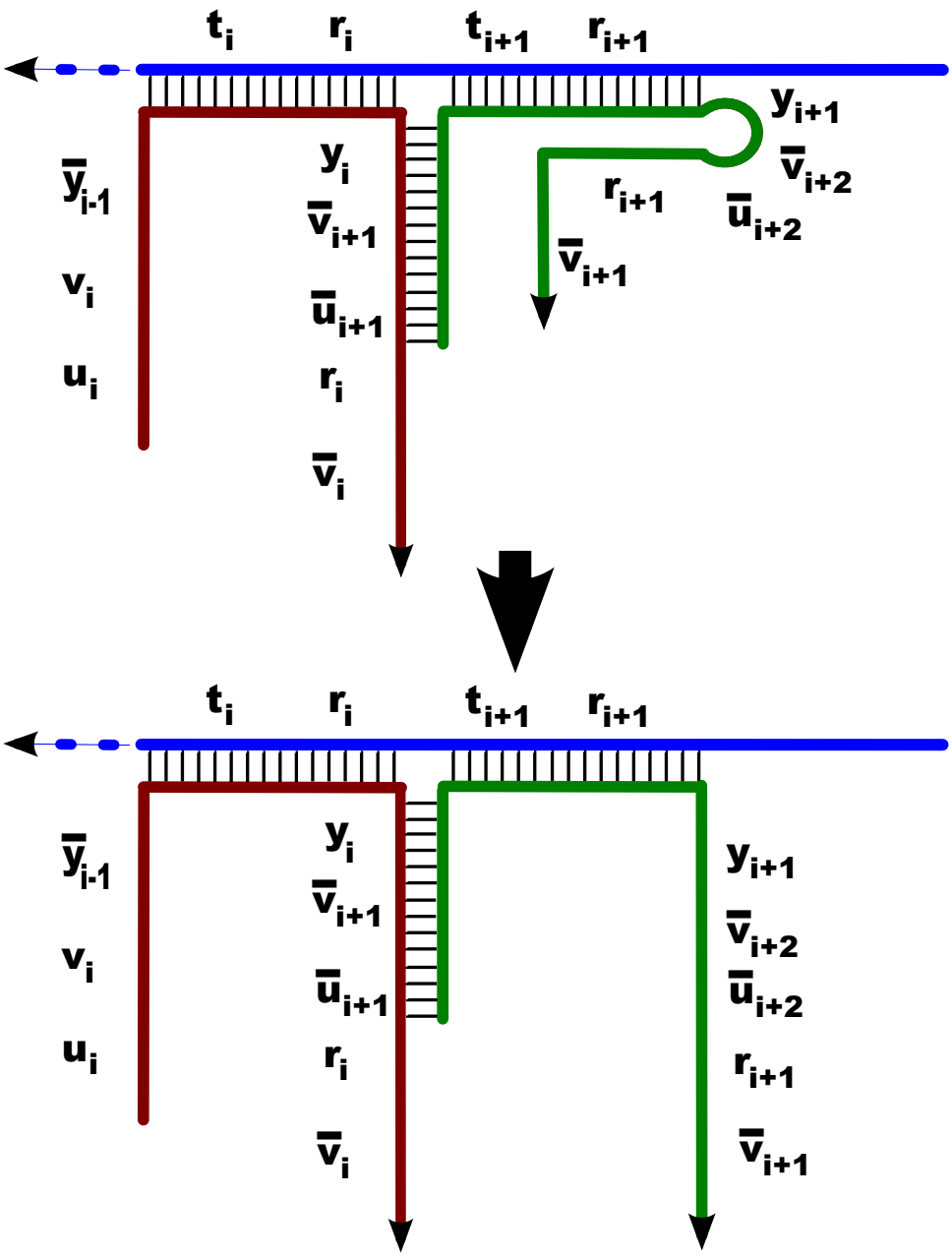

Figure 4: First protocol - III

3 Second Protocol for high-fidelity hybridization

As before, let $s = r_n t_n r_{n-1} t_{n-1} \ldots r_{i+1} t_{i+1} r_i t_i \ldots r_1 t_1$ be a long DNA target strand and $c_i : i = 1, \ldots, n$ be short checker sequences that bind to s. Figure 5 indicates two consecutive checker sequences c_i in red and c_{i+1} in green. The expected secondary structure of these strands is also indicated in the figure. As an inductive hypothesis, assume that c_i is bound to the blue template strand s as shown in Figure 5. We claim that the appropriate complementary portion of strand c_{i+1}, $\bar{r}_i \bar{t}_{i+1} \bar{r}_{i+1}$ binds to $r_{i+1} t_{i+1} r_i$ on s and the induction is advanced by one step. The chief difference between this protocol and the earlier one is that the incoming short strand first strand invades from $5'$ to $3'$ on the blue strand and then uses this toehold to strand invade the rest of the subsequence from $3\prime$ to $5'$.

First, u_{i+1} on the green strand binds to its complementary region $\bar{u}_{i+1}$ on the red strand (Fig. 5). Through this toehold, a strand displacement reaction occurs, breaking the bond between v_{i+1} and $\bar{v}_{i+1}$. v_{i+1} is now attached to its complementary region $\bar{v}_{i+1}$ on the red strand. Now the hairpin structure r_i and $\bar{r}_i$ on the green strand can invade from $5'$ to $3'$ on the blue target strand s while at the same time attaching to $\bar{r}_i$ on the red strand (Fig. 6). This breaks the bond between r_i on the blue target strand and $\bar{r}_i$ on the red strand. Now, a strand displacement reaction occurs via the toehold r_i on the blue target strand which opens a hairpin structure by breaking the bonds between $r_{i+1} t_{i+1}$ and $\bar{r}_{i+1} \bar{t}_{i+1}$ on the green strand, allowing $\bar{r}_{i+1} \bar{t}_{i+1}$ on the green strand to bind to $r_{i+1} t_{i+1}$ on the blue strand (Fig. 7). This opens the hairpin $\bar{v}_{i+2} \bar{u}_{i+2}$. Thus, the green strand is in the same conformation as the blue was at the beginning of the induction step. The sequence $\bar{v}_{i+2} \bar{u}_{i+2}$ on the green strand can now open up the hairpin on strand c_{i+2}, thus activating it and the process continues till all c_i are bound to the template s. If the potential target s does not have the appropriate sequence, some checker strand will not bind and hence the sequence of attachments will be halted.

4 Potential Applications of High-Fidelity DNA Hybridization

Our set of checker sequences can be thought of as a high-fidelity rationally programmed aptamer for a specific DNA target sequence. Our checker sequences can be extended into functional nanostructures that interact with the target sequence, for example as a molecular cage that encapsulates the target molecule. The completion of subsequence verification can trigger other reactions and prove useful in molecular detection.

There are many significant applications of our protocols for High-Fidelity DNA Hybridization. These include significantly increasing the specificity of :

- DNA enzymatic reactions such as restriction cuts
- Priming of PCR reactions used for amplification
- DNA hybridization arrays
- Binding of the pads of DNA tile nanostructures together to form DNA lattices
- Hybridization between long scaffold strands and short sticker strands
- DNA computation reactions involving DNA hybridizations (see section 4.1)

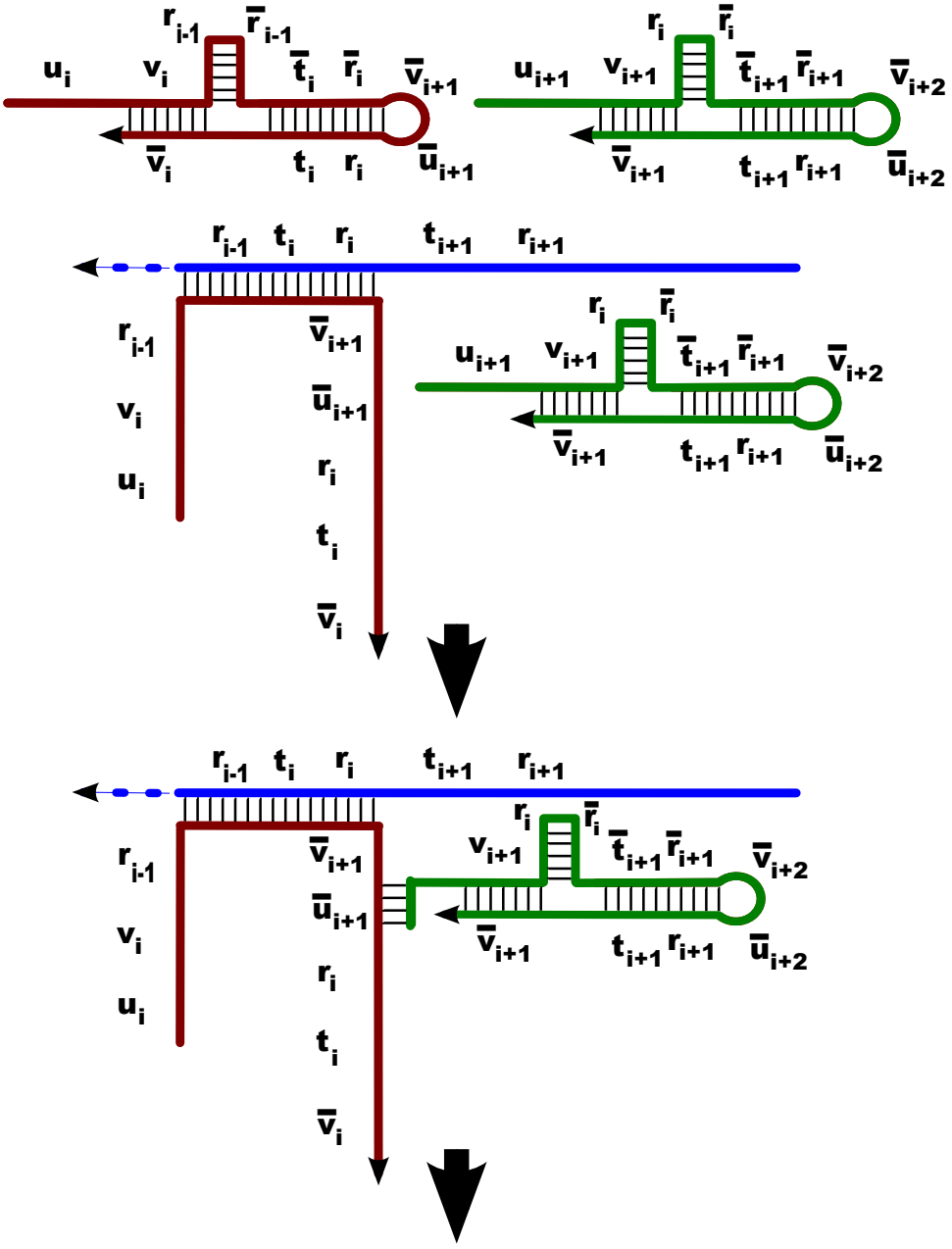

Figure 5: Second protocol - I

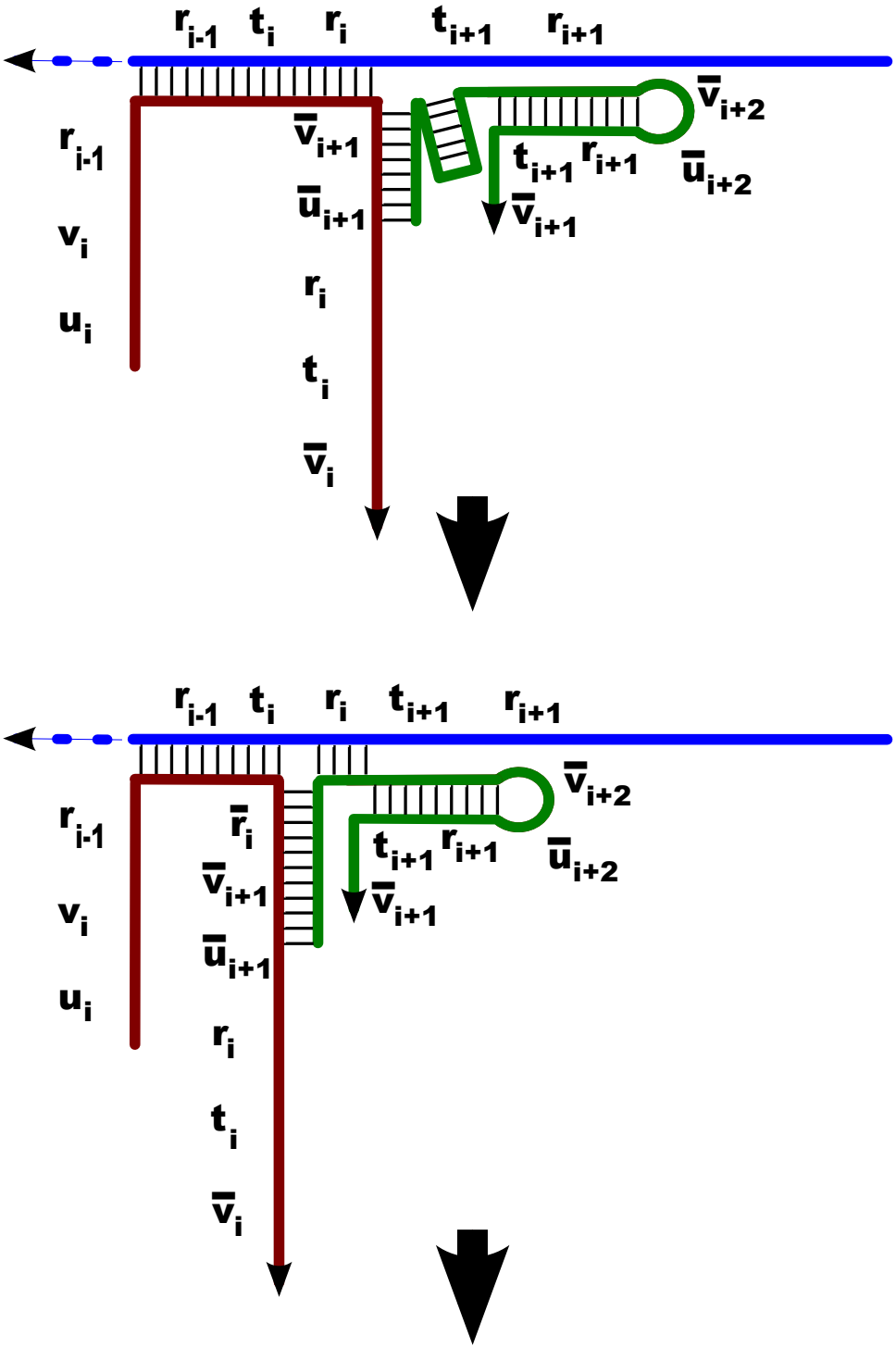

Figure 6: Second protocol - II

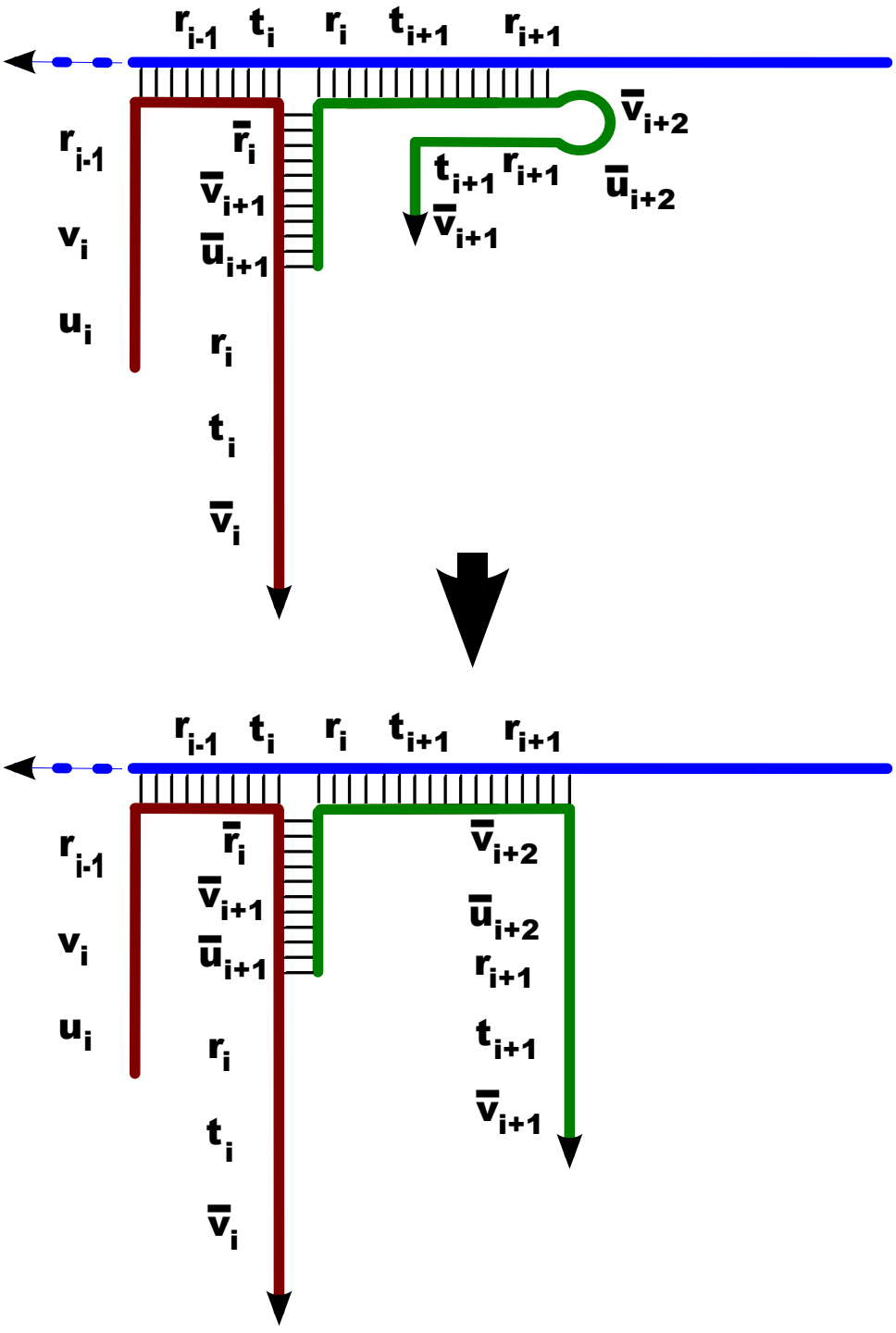

Figure 7: Second protocol - III

4.1 Simulation of Deterministic Finite Automata

$M = \{\Sigma, S, s, F, \delta\}$ is a deterministic finite automaton where Σ is the finite *input* alphabet, S is a finite set of *states*, $s \in S$ is the *start* state, $F \subseteq S$ is the set of *accepting* states and $\delta : S \times \Sigma \to S$ is the *transition* function. The *accepting* function of M, $\Delta_M : S \times \Sigma^* \to S$, is defined recursively for any $a \in S, \beta \in \Sigma, x \in \Sigma^*$ as $\Delta_M(a, \beta x) = \Delta_M(\delta(a, \beta), x)$ and $\Delta_M(a, \epsilon) = a$. M accepts the language $L_M = \{w \in \Sigma^* | \Delta_M(s, w) \in F\}$.

Our protocols for high-fidelity hybridization can be adapted to implement any deterministic finite automaton. For simplicity assume that the input alphabet is $\Sigma = \{0, 1\}$. We will adapt the protocol outlined in section 2 for our simulation. Let $w = \beta_1\beta_2 \ldots \beta_n$ be the input to the automaton, where $\beta_i \in \{0, 1\}$. We encode this input into the target strand along with an initiation sequence $\triangleright$ and a completion sequence $\triangleleft$ as $\triangleleft, \beta_n, \beta_{n-1}, \ldots, \beta_1, \triangleright$. In the notation used in section 2 the subsequence $r_i t_i$ encodes for the symbol β_i for $i = 1, 2, \ldots, n$. The protocol must check whether this input is in the language accepted by the automaton. The checker sequences will encode the transition function δ by associating a (state, symbol) combination with the appropriate next state. For example suppose the input sequence $\beta_1\beta_2 \ldots \beta_i$ has been consumed by the automaton and the current state is a. This current state will be encoded inside the i^{th} checker sequence as the subsequence $y_i \bar{v}_{i+1} \bar{u}_{i+1}$. The next transition $\delta(a, \beta_{i+1}) = b$ is executed by the $(i+1)^{\text{th}}$ checker sequence which contains the subsequences $u_{i+1} v_{i+1} \bar{y}_i$ (which codes for a), $\bar{t}_{i+1} \bar{r}_{i+1}$ (which codes for β_{i+1}) and $y_{i+1} \bar{v}_{i+2} \bar{u}_{i+2}$ (which codes for b) by hybridizing to complementary regions on the i^{th} checker sequence and the target. If the checker sequences successfully attach all the way to $r_n t_n$ this indicates that the target sequence encodes an input that should be accepted by the automaton. The special sequence $\triangleleft$ is initially in the form of a hairpin and the attachment of the last checker sequence opens the hairpin, which can be detected by fluoroscent emission due to the spatial decoupling of a fluorophore-quencher modified pair of DNA bases on the subsequence $\triangleleft$.

Note that instead of the correct $(i+1)^{\text{th}}$ checker sequence encoding the pair (a, β_{i+1}) an incorrect checker sequence that encodes $(a, \bar{\beta}_{i+1})$ may also attach to the previous checker sequence. However, it will not attach to the target and hence its second hairpin will remain intact, blocking further attachments. Thus, there is at least one of two choices of attachment at each step that further the process towards completion. If we assume equal probabilities of attachment for such competing checker sequences, an n length input has probability at least 2^{-n} of successful completion. This is a very low probability for long input sequences and alternate strategies to undo incorrect checker sequence attachment must be considered in this case. A key advantage of this simulation is that no matter what the input the same fixed set of checker sequences can check for acceptance of the input by the automaton. The number of such checker sequences is twice the number of states of the automaton. In fact, due to the fidelity of the hybridizations, multiple inputs may be processed in a parallel manner. By using fluorophores with distinct emission wavelengths on different targets we can detect multiple outputs in parallel. The simulation process is autonomous and does not use enzymes. It is easy to adapt this protocol to simulate non-deterministic finite automata and we leave the details to the reader.

5 Kinetic models for strand displacement

Hybridization fueled DNA strand displacement reactions can be used in a variety of nanoscale protocols to achieve molecular transportation (Sherman and Seeman (2004), Turberfield et al. (2003)), motors (Yurke et al. (2000)), detection (Dirks and Pierce (2004)) and computing (Zhang et al. (2007); Yin et al. (2005)). A thorough understanding of this fundamental process is key to the design of more complicated and involved dynamic DNA devices. Simple strand displacement pro-

cesses have been studied theoretically as one-dimensional random walks paving the way for simple stochastic models. A useful approach to understanding strand displacement is via kinetic computer simulations. Computer simulations can be a high-level tool for a cheap, quick and controlled investigation of the processes underlying strand displacement. Some of the information processing circuits of Zhang et al. (2007) that use simple strand displacement processes have been simulated (Phillips and Cardelli (2009)), but more general reactions involving multiple strands and strand exchange have not yet been studied. Involved DNA devices using complicated strand displacement reactions like the ones proposed in this paper for high-fidelity hybridization are a challenging and appropriate test case for such kinetic computer simulation studies. Experimental data about the kinetics of strand displacement is available (Green and Tibbetts (1981); Turberfield et al. (2003)) and can be used to create simulations that test the feasibility of the protocols proposed in this paper and improve their efficiency. These simulations can be thought of as a high-level testing and debugging environment that allow for improved designs of these involved protocols.

6 Conclusion

6.1 Experimental Verification

We propose a simple experiment using just two checker sequences to test if it can distinguish a target sequence with high-fidelity from an ensemble of moderately long sequences. We propose to use gel electrophoresis data and fluoroscence resonance energy transfer (FRET) to test for the fully-complemented target sequence. Let the sequence of the target strand be $s = r_2 t_2 r_1 t_1 \triangleright$ with each of the subsequences relatively short. $\triangleright$ is a special subsequence to initiate the process. We can introduce *noise* in the form of other strands with sequences differing from that of the target. The two checker sequences are c_1 and c_2, like in fig. 2 for the first protoocol or like in fig. 5 for the second protocol. There is an initiation sequence c_0 that binds to $\triangleright$ on the target and initiates binding of c_1. The terminal base at the $5'$ end of s can be modified with an emmitive fluorophore while the corresponding complementary base on the subsequence $\bar{r}_2$, part of c_2, can be modified with a corresponding quencher. If c_2 binds to the target s, the fluorophore-quencher pair will be brought in close proximity ($< 4nm$) and the fluoroscent emission of the fluorophore will be quenched signalling successful complete hybridization of the target. Appropriate controls can be devised for testing the specificity of binding of the checker sequences by modifying with an emmitive fluorophore the $5'$ ends of the *noisy* strands. In the absence of the target strand in the ensemble, no quenching of the fluoroscent signal should be observed.

6.2 Discussion

The problem of high-fidelity hybridization in long nucleic acid strands is a fundamental challenge with applications to a wide range of issues that arise frequently in biological nanotechnology. In this work, we have proposed two protocols for achieving highly specific hybridization to a specific target strand at the exclusion of all other strands in solution. We chiefly rely on hybridization between short segments of complementary DNA (which are inherently highly specific), strand displacement reactions and energy released by conversion of ssDNA to dsDNA to overcome kinetic traps in the form of meta-stable hairpins. We proposed simple experiments to test out our protocols in the base case where there are only two checker sequences that each verify half of the target strand. We wish to test whether our protocols will scale for much longer target sequences requiring many more checker sequences and hence we propose a computer based prediction of these protocols via

thorough simulations of strand displacement reactions and kinetic behaviour of kinetic energy traps in the form of meta-stable hairpins.

References

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

Green, C. and Tibbetts, C. (1981). Reassociation Rate Limited Displacement of DNA Strands by Branch Migration. *Nucleic Acids Research*, 9(8):1905–1918.

Phillips, A. and Cardelli, L. (2009). A Programming Language for Composable DNA Circuits. *Journal of The Royal Society Interface*, 6(11):419–436.

Sherman, W. and Seeman, N. (2004). A Precisely Controlled DNA Biped Walking Device. *Nano Letters*, 4:1203–1207.

Tian, Y., He, Y., , Chen, Y., Yin, P., and Mao, C. (2005). A DNAzyme That Walks Processively and Autonomously along a One-Dimensional Track. *Angewandte Chemie International Edition*, 44(28):4355–4358.

Turberfield, A., Mitchell, J., Yurke, B., Mills, A., Blakey, M., and Simmel, F. (2003). DNA Fuel for Free-Running Nanomachines. *Physical Review Letters*, 90(11).

Yin, P., Sahu, S., Turberfield, A., and Reif, J. (2005). Design of Autonomous DNA Cellular Automata. *DNA Computing*, pages 376–387.

Yin, P., Yan, H., Daniell, X., Turberfield, A., and Reif, J. (2004). A Unidirectional DNA Walker Moving Autonomously Along a Linear Track. *Angewandte Chemie International Edition*, 116(37):5014–5019.

Yurke, B., Turberfield, A., Mills, A., Simmel, F., and Neumann, J. (2000). A DNA-fuelled Molecular Machine Made of DNA. *Nature*, 406(6796):605–608.

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