LUND UNIVERSITY

Design and Construction of a One- Dimensional DNA Track for an Artificial Molecular Motor

Kovacic, Suzana; Samii, Laleh; Woolfson, Derek N.; Curmi, Paul M. G.; Linke, Heiner; Forde, Nancy R.; Blab, Gerhard A.

Link to publication

Research Article

Design and Construction of a One-Dimensional DNA Track for an Artificial Molecular Motor

Suzana Kovacic,[1] Laleh Samii,[1] Derek N. Woolfson,[2, 3] Paul M. G. Curmi,[4, 5] Heiner Linke,[6] Nancy R. Forde,[1] and Gerhard A. Blab[7]

[1] *Department of Physics, Simon Fraser University, 8888 University Drive, Burnaby, BC, Canada V5A 1S6*
[2] *School of Chemistry, University of Bristol, Bristol BS8 1TS, UK*
[3] *School of Biochemistry, University of Bristol, Bristol BS8 1TD, UK*
[4] *School of Physics, University of New South Wales, Sydney, NSW 2052, Australia*
[5] *Centre for Applied Medical Research, St. Vincent's Hospital, Darlinghurst, Sydney, NSW 2010, Australia*
[6] *The Nanometer Structure Consortium (nmC@LU) and Division of Solid State Physics, Lund University, Box 118, 22100 Lund, Sweden*
[7] *Molecular Biophysics, Utrecht University, Postbus 80'000, 3508 TA Utrecht, The Netherlands*

Correspondence should be addressed to Gerhard A. Blab, g.a.blab@uu.nl

DNA is a versatile heteropolymer that shows great potential as a building block for a diverse array of nanostructures. We present here a solution to the problem of designing and synthesizing a DNA-based nanostructure that will serve as the track along which an artificial molecular motor processes. This one-dimensional DNA track exhibits periodically repeating elements that provide specific binding sites for the molecular motor. Besides these binding elements, additional sequences are necessary to label specific regions within the DNA track and to facilitate track construction. Designing an ideal DNA track sequence presents a particular challenge because of the many variable elements that greatly expand the number of potential sequences from which the ideal sequence must be chosen. In order to find a suitable DNA sequence, we have adapted a genetic algorithm which is well suited for a large but sparse search space. This algorithm readily identifies long DNA sequences that include all the necessary elements to both facilitate DNA track construction and to present appropriate binding sites for the molecular motor. We have successfully experimentally incorporated the sequence identified by the algorithm into a long DNA track meeting the criteria for observation of the molecular motor's activity.

1. Introduction

DNA presents significant advantages as a building block for nanostructured materials, as specific sequences of DNA can be easily manipulated using a large set of biochemical tools and protocols. These can both modify its existing biological functions and design and construct completely novel functionality [1]. In addition to its standard role of encoding genetic information, DNA also possesses unique physical properties that make it amenable for use as a template or material for nanostructures. These include its mechanical stiffness [2] and its ability to self-assemble driven by the specific recognition of complementary bases. Exciting new applications have been designed which exploit the ease of use of DNA. "Nonstandard" DNAs include such examples as DNA origami, in which the directed folding into arbitrary 2- and 3-dimensional shapes is achieved by DNA crosslinks or "staples" [3]. This technique promises immense potential to use DNA as a building block for self-assembled nanopatterning [4] or for presenting an array of substrates for addressable chemical reactions on the single molecule level [5]. Other examples include "DNA spiders," which have been designed to exhibit molecular motor properties [6–9], active folds of DNA called aptamers [10], which can be used as highly sensitive biosensors, and catalytic nucleic acid structures, ribozymes [11]. Another type of structures

is based on the self-recognition of the DNA base guanine that can generate linear aggregates, termed "G-wires," which have shown promise as the basis of future molecular-scaled electronic devices [12]. Despite the many different designs and applications of these nonstandard DNAs, one common challenge for these uses of DNA is the task of identifying and linking biologically relevant motifs into functional units [13]. This challenge becomes increasingly more complex the further away one moves from an existing biological system to a man-made structure.

In this paper, we report on the design and construction of a periodic one-dimensional DNA track for a novel artificial molecular motor, the tumbleweed (TW) [14, 15]. The tumbleweed is designed to consist of a central hub connecting three different DNA repressor proteins: R_A (methionine repressor MetJ (Q44K) [16, 17]), R_B (tryptophan repressor TrpR [18, 19]), and R_C (purine repressor PurR [20, 21]). Each repressor acts as a foot of the motor that can be made to bind and unbind from its well-defined recognition sequence on the DNA track, depending on the concentration of its specific ligand present in solution [22–24]. The directionality of stepping is determined both by the order of ligand exchanges and by the spatial order of the recognition sequences A, B, and C on the periodic track. We propose to implement a DNA curtain technique [25] to allow simultaneous measurements of the movement of many TW motors along parallel DNA tracks (Figure 1).

On the length scale of the TW step, DNA is not a featureless molecule but presents two distinct sides (grooves) that twist helically around the length of the molecule with an 11 base-pair (bp) pitch. Therefore, one needs to ensure that all three recognition sites present on the same side of the DNA molecule to avoid steric hindrance caused by the motor "corkscrewing" around the DNA track instead of stepping along it. This is especially a concern if the track is on or close to a surface as is typically the case for single-molecule tracking experiments [26].

The proper sequence, spacing, and spatial orientation of the recognition sites on the track are thus important requirements for the design of the DNA track. In addition, in order to demonstrate motor motion over significant distances, the track needs to be highly repetitive. Furthermore, we need to be able to modify an end of the DNA construct so that it can be anchored and stretched on a surface and form a linear track along which we can observe the movements of the TW motor. A further challenge is presented by the unavoidable drift of the sample on the microscope during the single-molecule visualization of tumbleweed dynamics. As such, the DNA track needs to incorporate fluorescent fiducial markers at specified intervals that provide a readout of this drift, allowing us to measure the position of the TW relative to the track, thus removing any effect of the movement of the whole sample during the experiment. Lastly, the track has to be designed in a way so that it can be constructed and modified easily and quickly.

Here we present a method for constructing a long DNA track containing a periodic sequence of repressor binding sites. The problem of design is both theoretical—how to find
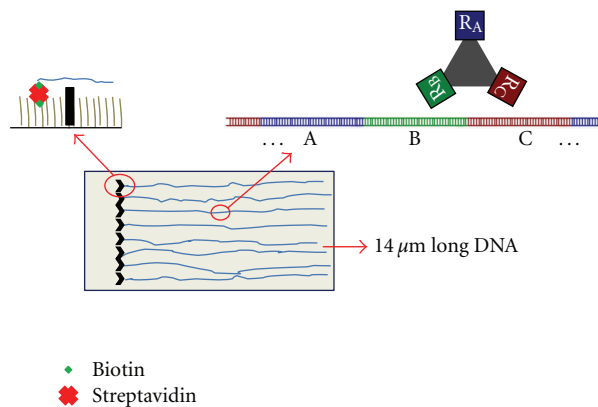

Figure 1: Schematic of the DNA track and the TW motor running along it. The track is anchored to a surface, which has been made inert by means of a lipid bilayer, using a biotin/streptavidin interaction (left top). Several DNA tracks of about $14\,\mu m$ length can be stretched into a DNA curtain and allow observation of the movement of many TW motors at the same time. The DNA track consists of several repeats of the ABC motif of recognition sequences along which the tripedal TW motor steps (top right: R_A, R_B, and R_C indicate the three repressor "feet," step size ~11 nm).

a sequence with the desired properties—and experimental—how to quickly and economically construct a long, periodic sequence of DNA. This design problem has relevance not only to molecular motor experiments but also to many of the aforementioned applications. Our experimental approach takes advantage of modular "cassette" sequences of the form $L(ABC)_N R$. Each cassette contains N-repeating units of the ordered recognition sequences A, B, and C. We show how cassettes with large N can be generated from a primary cassette with $N = 1$, by designing well-defined handles on both ends (L, R) of the cassette. The handles also allow us to specifically bind one or both ends of the track to a surface, or to attach the fiducial markers that will be necessary to follow the progress of the TW motor along the track. We, furthermore, describe the development and application of a genetic algorithm used to select an appropriate sequence for this primary cassette, which satisfies the imposed constraints. Finally, we illustrate the incorporation of the sequence selected by the algorithm into a long (>10 μm) DNA track as required for DNA curtains, suitable for visualization experiments of motor dynamics.

2. Results and Discussion

2.1. Approach. The tumbleweed (TW) motor is designed to be a self-assembled biomolecular complex that can move by rectified diffusion along a DNA track. It consists of a triangular, peptide-based hub connecting three protein-"feet." The hub can be labelled with a fluorophore, a quantum dot, or a small colloidal bead to facilitate detection. Each foot of the TW motor is a DNA-repressor protein, approximately 5 nm in size, that can attach tightly to its DNA recognition sequence in the presence of a specific ligand molecule, and which will release quickly once the ligand

is removed. The feet are attached to the hub by means of flexible linkers or "legs," such that the motor can reach the next recognition site on the track and take steps of approximately 11 nm in size (Figure 1).

The DNA track provides a one-dimensional substrate along which the TW motor processes by presenting multiple repeats of the A, B, and C recognition sites in the appropriate sequence to allow binding and unbinding of the repressor feet in the presence of the appropriate ligands. The track takes into consideration such factors as appropriate spacing and orientation such that the TW motor steps along one face of the DNA molecule rather than twisting around it. To establish the necessary spacing between the repressor sequences in this ABC unit, models of all three repressors (from crystal structures of the repressors complexed with DNA, namely, PDB structures 1TRO, 1MJM, and 1F5T) were docked to a model of ideal B-form DNA (nucleic acid builder NAB), by performing least squares fits involving all atoms in the DNA portion of the crystal structure. Starting from equal spacing between the binding sites, the repressor models were moved until all resided on the same side of the DNA track. Additional sites are necessary to allow the incorporation of fiducial markers and to provide attachment points by which the DNA molecule can be bound and stretched into a linear track.

2.2. DNA Track Design by Genetic Algorithm. Our first challenge in this project was to develop the necessary tools to find a suitable DNA sequence that incorporates the appropriate repressor binding sequences with the correct spacing and orientation as well as additional sequences necessary for construction of the DNA track. For reasons of flexibility and economy, we decided to follow a bottom-up process in which we built the primary cassette *de novo* from oligonucleotides. The 11 nm stepsize of the TW motor translates into a repeating ABC unit which is 103 bp long. Together with the flanking sequences, which are required for the anchoring, doubling, and amplification of the track, the primary cassette has a total length of 169 bp (Figure 2). 101 bp of the cassette are fully defined by being in a recognition, restriction, or connection site, and 58 are variable, that is, only partly defined or arbitrary. However, only a small subpopulation ($p \ll 10^{-4}$) of the resulting 2.60×10^{33} distinct sequences is actually usable for our purpose. Besides the obvious requirement that no additional recognition or restriction sites may be formed by the variable bases, the different parts of the resulting cassette also need to present amenable thermodynamic properties (in our case the melting temperature), be easily constructible from oligonucleotides, and need to be sufficiently dissimilar to each other to minimize deletion and recombination artifacts in later molecular biological steps. An attempt to find a suitable primary cassette by brute force—testing several thousand sequences by randomly selecting the variable bases—did not yield satisfactory results. Unsurprisingly, such a brute force search is feasible only for very short cassettes, as some of the required tests (such as hybridization tests and calculation of the melting temperature) scale super-linearly with sequence length. In consequence, we describe here our implementation of a genetic algorithm [27, 28] to find an acceptable cassette within the search space that meets all the requirements to synthesize the desired DNA track.

Our approach is centered on the design of a primary cassette $L(ABC)_1R$, also called K_1, as shown in Figure 2 where A, B, and C correspond to the repressor binding site sequences and L and R correspond to the left and right handles which facilitate ligation and cassette doubling. An extra sequence, marked as "internal primer," was designed to fit between C and A and contains additional recognition sites, for example, for a fiducial marker. This sequence was not used in the initial experiments.

In order to successfully identify a DNA sequence, a search algorithm is needed that can be employed practically on a large, sparse search space. Appropriately, genetic algorithms are exceedingly well suited for this type of problem [27–29]. We implemented the algorithm in Octave [30]. The scripts, which are compatible with the commercial MATLAB platform, are available from the authors. The thermodynamic properties of DNA are calculated using the freely available UNAFold suite of programs by Markham and Zuker [31, 32]. In a genetic algorithm, initially random sequences are repeatedly evaluated (scored) according to predetermined properties and restrictions and allowed to recombine according to their score. In our example, the score was the sum of penalties incurred by a sequence. A "high" penalty of 0.05/base was assessed for bases that resided in user-defined sequences exhibiting unsatisfactory thermodynamic properties—that is a melting temperature that deviated $10°C$ or more from the target melting temperature ($56°C$)—or formed an unwanted restricted or recognition site. If the melting temperature of a subsequence deviated from the target by between $5°C$ and $10°C$, or if the same base occurred more than three times in a row, a lower penalty of 0.01/base was used. Additionally, each sequence had a small chance to incur random changes (mutations) in every round. As only the best scored sequences "survive" between rounds, the sequences quickly converge on a solution.

The parameters of the algorithm, that is, the scoring function (see Section 4), the number of sequences in the gene pool (100), the rate of random sequence changes ("mutations," $p_m = 10^{-3}$/base), and the magnitude of the penalties, were initially optimized and are now robust. The goal was to set the penalties and mutation rate low enough to ensure, on average, one or less random change per 169 bp sequence, but to make the gene pool large enough to obtain at least one random change per generation in the population. Small populations, or populations with too small penalties, tend to get stuck with nonoptimal results, while too large mutation rate p_m or penalties that are too high will not allow the population to retain good sequences. Figure 3 shows a screen shot of the genetic algorithm during its run. Each line in the graph represents one candidate sequence, and the colour indicates the penalty, that is, the deviation from a practical cassette, on a base-pair by base-pair level. Beneficial properties quickly spread in the population (the leftmost handle has a low or no penalty for all candidates). The algorithm usually terminates within 30 cycles and will find a slightly different cassette every time it is run.
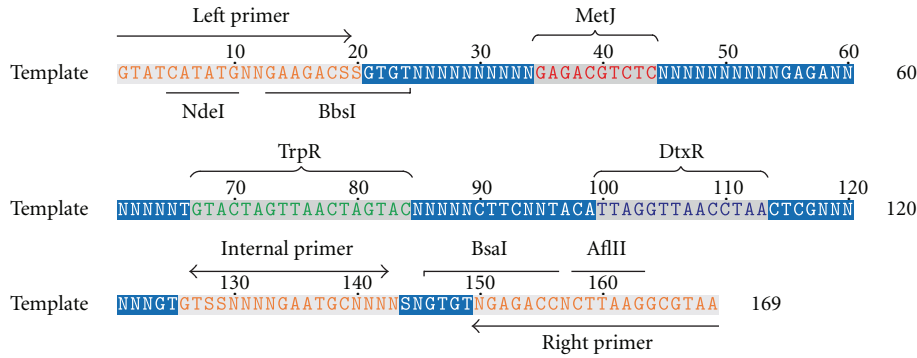
FIGURE 2: The 169 base-pair design template for a cassette $L(ABC)_1R \equiv K_1$ for building a motor track. The template contains an additional sequence, marked "internal primer," which would allow for modification, for example, to attach a fiducial marker. Necessary design elements are noted: brackets mark repressor recognition sequences; arrows represent potential primer sites (also referred to as "handles" in the text); bars show where the template can be cut using restriction enzymes. In addition to the fully determinate bases (ATCG), the template also contains ambiguous bases (N for "any" base, S for a strong base, i.e., C or G). This template describes a total of 2.60×10^{33} distinct sequences.
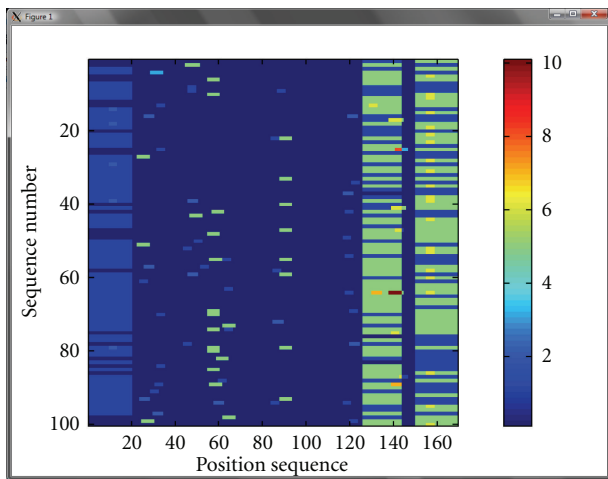

FIGURE 3: Screen shot of the genetic algorithm during its search for a suitable sequence for the primary cassette K_1. Each line represents one candidate sequence of 169 bp, and colour denotes penalty scores assessed.

2.3. DNA Track Construction. The algorithm yielded a number of suitable sequences that could be employed to construct the desired DNA track. Table 1 shows the 10 oligonucleotide sequences chosen for the actual construction of the DNA track. Every pair of complementary oligonucleotides forms nonpalindromic single-stranded overhangs that allow it to be ligated to its neighbours. The primary cassette L(ABC)R, which is also referred to as K_1, was generated by annealing and sequential ligation as described in Section 4 ("Assembly and Amplification of the Primary Cassette," below) and K_1 is 142 bp long and contains a single ABC unit where A, B, and C correspond to the repressor binding sites for MetJ, TrpR, and DtxR, respectively. The primary cassette K_1 was subsequently subcloned into a high copy-number bacterial plasmid (pYIC), thus, creating a primary plasmid pK_1. Working with a plasmid permits the production of

sufficient quantities of DNA, and thus the track itself, by growing it in bacteria. The sequences of the handles L and R were taken from the plasmid itself, to permit easy insertion and subsequent handling.

To observe the movement of the TW motor, the track must be sufficiently long such that it contains multiple copies of the ABC repeat along which the TW motor can move. In order to construct the long track, we made use of the L and R handles flanking the ABC repeat which each carry important restriction sites. A digest of the plasmid pK_N (where N stands for the number of ABC repeats in the plasmid) using the restriction enzymes NdeI and BbsI linearizes the plasmid and excises a short, unneeded 14 base-pair region in the left handle. Digesting a pK_N plasmid instead with NdeI and BsaI excises the K_N cassette. This can then be ligated into the linearized plasmid to form a new, circular plasmid DNA containing a doubling of the ABC unit (Figure 4). The localization of the enzyme recognition and restriction sites are shown in Figure 2. After each doubling, the new plasmid was directly subcloned into the chemically competent *E. coli* strain DH5α for amplification (see Section 4). We have taken care to design the template such that the restriction sites are maintained exclusively at the ends of the cassette in this doubling process so that this step can be repeated. Thus, starting from pK_1, after only 5 doubling steps, we can go from a 30 nm $(ABC)_1$ track to a plasmid pK_{32} which contains 32 copies of the ABC repeat and measures over 1 μm.

One difficulty raised by this doubling scheme is the repetitiveness of the track itself. Bacteria do not usually tolerate repetitive sequences and tend to remove them from plasmids rather efficiently [33]. The decision to choose DH5α was based on its tolerance for DNA repeats due to its lack of the recA protein that is responsible for recombination of homologous (repeating) sequences of DNA. Figure 5 shows plasmids ranging from pK_1 to pK_{32} double-digested with BbsI and BsaI. These restriction enzymes have their recognition sites in the left and right handle regions and thus excise the $(ABC)_N$ cassette as a whole from the plasmid (and also generate two fragments of constant length 1221

TABLE 1: Sequence of the individual oligonucleotides used to construct the repetitive TW track; the top oligo of each pair is shown 5′ → 3′, while the complementary strand is shown in reverse 3′ → 5′ direction. Bold bases indicate the recognition sites for the TW protein feet. This primary cassette K_1 is shorter than 169 bp because it does not contain the sequence for the fiducial marker shown in Figure 2.

| Handle L | GTATCATATGCTGAAGACGG |
	CATAGTATACGACTTCTGCCCACA
A	GTGTCATATAGCCA**GAGACGTCTC**GGCACGCGTC
	GTATATCGGT**CTCTGCAGAG**CCGTGCGCAGCTCT
B	GAGAATTTAAAT**GTACTAGTTAACTAGTAC**GGTGA
	TAAATTTA**CATGATCAATTGATCATG**CCACTGAAG
C	CTTCAGCTCC**TTAGGTTAACCTAA**CTCGAGCATA
	TCGAGG**AATCCAATTGGATT**GAGCTCGTATCACA
Handle R	GTGTAGAGACCACTTAAGGCGTAA
	TCTCTGGTGAATTCCGCATT
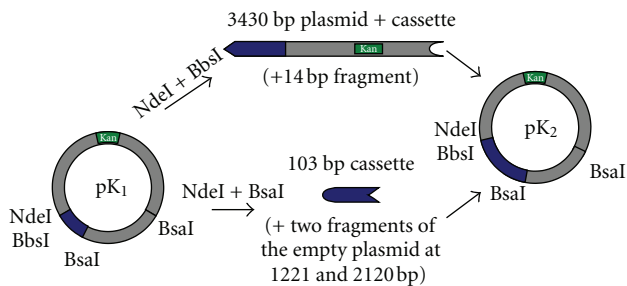
FIGURE 4: To double the length of the (ABC) unit, the pK_1 plasmid (left, with relevant restriction sites noted) was double-digested with either BbsI and NdeI (donor, including the kanamycin resistance gene "Kan," center top) or NdeI and BsaI (recipient, center bottom). Digestion with BbsI or BsaI produces compatible overhangs between donor and recipient. Donor and recipient fragments, obtained from gel purification of each double digest, were ligated together to form a pK_2 plasmid. The two repeats of the ABC sequence in the plasmid pK_2 are joined without forming additional restriction sites.
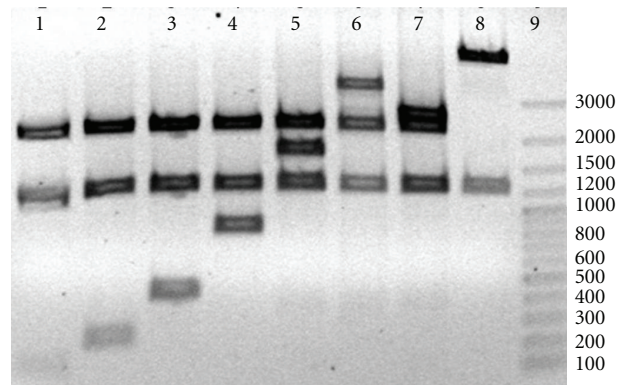

FIGURE 5: Agarose gel electrophoresis of BbsI/BsaI double-digested plasmids containing repeat sequences $(ABC)_N$. Lanes 1 to 6 correspond to $N = 1, 2, 4, 8, 16$, and 32. Lane 7 shows an unexpected $N = 24$ clone generated from a doubling of pK_{16}. Lanes 8 and 9 are controls (digest of the parent pYIC plasmid and 1 kbp ladder, resp.). Numbers on the right indicate fragment lengths in base pairs.

and 2120 bp). The cassettes show up as bands of increasing molecular weight in lane 1 to 6 of Figure 5. The cassette in lane 7, which was expected to be pK_{32}, appears shortened and demonstrates that highly repetitive constructs can become unstable even in DH5α, highlighting the need for vigilant monitoring of the constructs at each doubling step.

After obtaining a sufficiently long $(ABC)_N$ cassette in a plasmid, two steps remain to obtain a proper DNA track for our TW motor: modification of the handles to allow us to anchor the track to a surface and the attachment of fiducial markers.

For our initial experiments, we utilized the pK_8 plasmid, as it was determined to be stable in our chosen bacterial system and at the same time contained enough repeats of the ABC unit to provide a sufficient number of binding and unbinding steps by the TW motor. We propose to use DNA curtains [25] to stretch many DNA molecules in parallel and observe movement of TW motors along the tracks (Figure 1). The K_8 motif, however, is too short to be stretched as a DNA curtain, as this typically requires DNA which is more than $10\,\mu m$ long. We, therefore, lengthened the K_8 track from 250 nm to $14\,\mu m$ by attaching long DNA handles derived

from lambda DNA to the ends of the track. Lambda DNA is approximately $16\,\mu m$ long, contains a number of restriction sites which facilitate the insertion of the TW DNA track, and, importantly, does not contain repressor binding sites which could compete for binding with the TW motor. DNA curtains technology additionally requires a biotin tag at one end of the DNA molecule. Biotin is used to attach the DNA to a lipid bilayer via streptavidin (Figure 1; see Section 4.3 for details). The success of the ligation and end modification by biotinylation using a labelled primer was confirmed by agarose gel electrophoresis and blotting (Figure 6).

The TW motor itself will be labelled to visualize the motor dynamics. However, unless the track is also fluorescently labelled, it would be difficult to establish that the TW motor is stepping along the track rather than simply appearing to move because of drift of the sample. Fluorescent markers are thus incorporated into the DNA track to flank the $L(ABC)_N R$ motif. These act as fiducial markers to provide fixed reference points against which the movement of the TW motor is monitored. Here, we design a fiducial marker consisting of a highly fluorescent 300 bp region of DNA. Fiducial markers flanking a 32-repeat
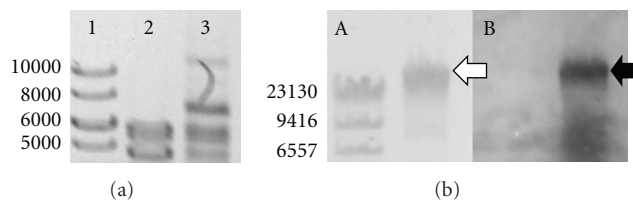
(a)   (b)

Figure 6: (a) EcoRV digestion of the K_8-lambda fragment ligation product shows that the ligation was successful, as evidenced by the extra high-molecular weight band in lane 3 compared with the digest of lambda DNA alone (lane 2). Lane 1 contains a 1 kb ladder with numbers on the left indicating fragment lengths in base pairs. (b) Biotinylation of the K_8-lambda DNA construct, using short biotinylated duplex DNA, was successful. (A) EtBr-stained gel showing the expected molecular weight of the construct (open arrow). (B) Construct transferred to a nitrocellulose membrane, probed with streptavidin horseradish peroxidase, and visualized by chemiluminescence (ECL Plus, Pierce). Streptavidin binding coincides with the predicted size of the construct (filled arrow), indicating the presence of a biotin label on this high-molecular-weight DNA.

region of ABC sequences would be separated by 3296 base pairs (approximately $1.1\,\mu$m). The use of a shorter ABC repeat sequence would result in fiducial marker sites that are closer together and are therefore difficult to resolve by fluorescence microscopy. The preliminary track built on the $(ABC)_8$ repeat sequence has the fiducial markers incorporated within the Lambda DNA handles to ensure resolution of the markers.

The fiducial marker is generated by PCR and subsequent click chemistry to incorporate fluorophores. The natural deoxythymidine triphosphate is fully replaced with an alkyne-modified deoxyuridine triphosphate during PCR amplification to generate a 300 bp DNA product incorporating 151 alkyne groups. The primers used to generate the 300-base pair DNA products were designed such that alkyne-modified deoxynucleotides are not incorporated into the restriction sequences which flank the two ends of the DNA. It is possible that alkyne modification at the restriction site could affect digestion and/or ligation. As such, EagI and PspOMI (recognition sequences—EagI CGGCCG; PspOMI GGGCCC) were chosen since these restriction sequences lack alkyne-modified bases following PCR amplification. Furthermore, these two restriction enzymes produce compatible sticky ends allowing end-to-end ligation of the 300-base pair pieces should a longer fiducial marker sequence be necessary. They also permit ligation into existing restriction sites flanking the $(ABC)_N$ cassette. Fluorescent DNA is generated by click chemistry between the alkyne-labelled DNA and azide-labelled carboxyrhodamine-110. When the fluorophore was added in 10-fold excess relative to alkyne (1500 equivalents per DNA molecule), the resulting highly fluorescent DNA exhibited unusual physicochemical properties as previously observed [31]. The DNA did not stain with ethidium bromide nor migrate as expected by agarose gel electrophoresis. When the amount of fluorophore was reduced to 10 mole equivalents per mole equivalent
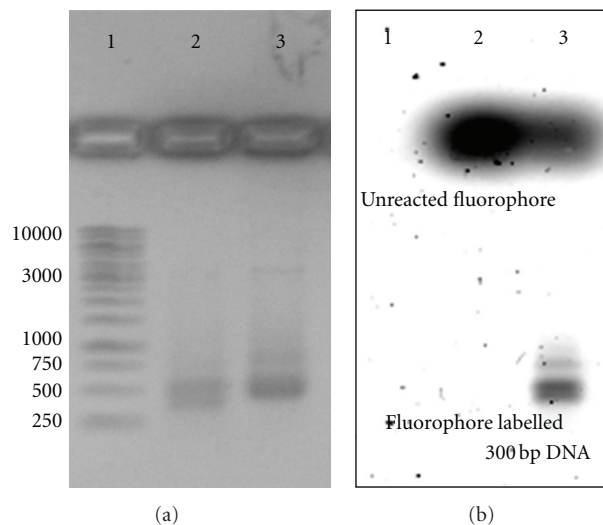

(a)   (b)

Figure 7: Labelling of 1 mole equivalent 300 bp alkyne DNA with 10 mole equivalents of carboxyrhodamine 110. A 1% agarose gel was stained with ethidium bromide (a) and scanned at 532 nm excitation and 555/20 nm emission (b) showing the reaction product between carboxyrhodamine 110-PEG3-azide and 300 bp alkyne-labelled DNA (lane 3) or 300 bp DNA (no alkyne incorporation) (lane 2). Lane 1: 1 kbp DNA ladder, numbers on the left indicating selected fragment lengths in base pairs.

DNA, the fluorophore-modified DNA migrated as expected on an agarose gel and could be detected by both ethidium bromide staining and fluorescence. Alkyne-labelled DNA ran slightly slower on a 1% agarose gel relative to control samples (Figure 7).

It is possible that fluorophore-modified DNA will be more difficult to incorporate at fiducial marker sites than alkyne-modified DNA due to the greater steric bulk of the fluorophore which may hinder digestion and/or ligation. Should this prove the case, alkyne-modified DNA will be ligated into the fiducial marker sites followed by click chemistry to introduce the fluorophore label. Alkyne-modified DNA has been successfully ligated in tandem demonstrating that the restriction enzymes EagI and PspOMI and T4 DNA ligase retain the ability to digest and ligate alkyne-modified DNA (Figure 8).

3. Conclusions

We have presented the *de novo* design and implementation of an extended one-dimensional DNA track for use in single-molecule experiments. For this goal, we have adapted molecular biology protocols which have allowed us to modularly construct, amplify, and extend a functional repetitive DNA track to a length of several micrometres. This track is amenable for modification and contains an anchoring point to the surface and fluorescent fiducial markers. Even though we introduce a specific application, the methods described in this paper are widely applicable to the construction of custom-made one-dimensional nanostructures based on DNA.
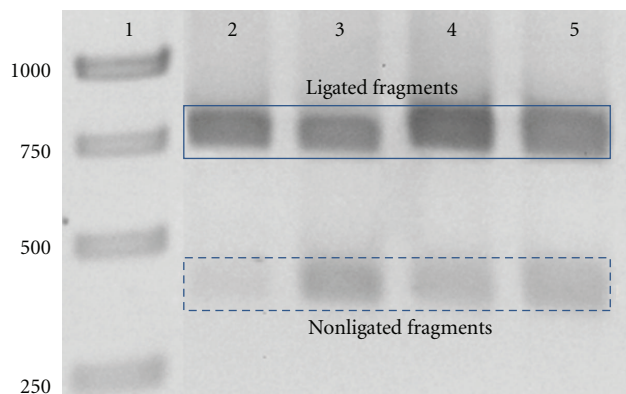
FIGURE 8: Alkyne modification does not interfere with digestion and ligation reactions. Lanes 4 and 5 show the successful self-ligation of 371 bp alkyne DNA after an EagI digest and a PspOMI digest, respectively, which each generate 738 bp fragments. Self-ligations of EagI- and PspOMI-digested control DNA (no alkyne modification) are shown in lanes 2 and 3, respectively. Lane 1 contains a 1 kb ladder, with numbers on the left indicating selected fragment lengths in base pairs. Alkyne DNA runs slightly slower than unmodified DNA.

At the heart of the paper is a computational toolbox that uses a heuristic search program, based on a genetic algorithm, to quickly and reliably identify suitable long DNA sequences with a set of specific structural, molecular biological, and thermodynamic properties from within the exceedingly vast, sparse space of possible DNA sequences. Our results show that by artful combination of established molecular biology techniques and targeted design using appropriate algorithms, completely new nanostructures with novel, not necessarily biological functions can be constructed from this versatile polymer.

4. Methods

4.1. Design of the Primary DNA Cassette. The genetic algorithm was implemented as a set of scripts in the freely available mathematical environment Octave [34] that is compatible with MATLAB (Mathwork, Natick, MA, USA). Thermodynamic properties of the DNA sequences of interest were established using an external application, UNAFold [31, 32].

At the heart of the genetic algorithm lies a template string that describes the general layout of the required DNA sequence using the upper-case one-letter codes. The allowed alphabet contains the standard bases "A," "T," "G," and "C", the IUBMB definitions for incompletely specified nucleobases [35], and the additional character "Q" indicating a nonbinding base. The template is divided into immutable "sequence features," which include defined base sequences such as recognition or restriction sites, and "subsequences," for which thermodynamic properties (entropic and enthalpic contributions to the binding energy, formation of secondary structures, or melting temperature T_M) can be defined. Subsequences can also be required to be unique, in which case the subsequence is tested for "substantial similarity" by calculating the relative difference in melting temperatures between the subsequence with its complement and with the rest of the template. A set of "restricted sites" (recognition and restriction sites) can be defined that must not occur anywhere but in predefined positions on either strand of the template. Restricted sites are again defined using the same upper-case one-letter codes for nucleobases as in the case of the template. The use of regular expressions [36, 37] is permissible.

The genetic algorithm starts with a set of n candidate sequences (user defined; typically $n \sim 100$) that conform to the sequence features specified for the template (see Figure 2 for the requirements of our template). Each candidate sequence is scored by applying penalties to bases that reside in subsequences exhibiting unsatisfactory thermodynamic properties (a melting temperature T_M outside a set range, or which are too similar to other parts of the track), form additional restriction or recognition sites, or are repetitive (Figure 3). The penalties are user defined and are typically graduated (high penalty, low penalty), to generate gradual convergence towards a satisfying sequence. Sequences in the lowest quartile survive and all sequences with higher penalty scores are deleted. The surviving sequences have then a small probability of mutation, proportional to the penalty assessed, typically at most 1 mutation per template, and are used to generate a new set of n candidates by random crossover. The algorithm terminates if a sequence without penalties is found, or if a set number of iterations (typically 50) is exceeded. The sequence chosen by the genetic algorithm and used in our experiments is shown in Table 1.

The collection of scripts is freely available from the authors.

4.2. Construction of the Primary Cassette

4.2.1. Assembly and Amplification of the Primary Cassette. The TW track was constructed modularly from a set of 10 oligonucleotides (Table 1: integrated DNA technologies). Each oligonucleotide was phosphorylated using T4 polynucleotide kinase (New England Biotech, NEB, Ipswich, MA, USA), following the manufacturer's protocol. Complementary oligonucleotides were annealed by heating to 85°C, followed by slow cooling ($\Delta T = 0.6°C/min$) to room temperature to generate five double-stranded DNA components. Three of these components correspond to the binding sites of the MetJ, TrpR, and PurR repressor proteins used in the TW motor and are denoted A, B, and C. The remaining two DNA components consist of sequences which permit subcloning and cassette doubling and are labelled L and R. The five components were ligated in the appropriate order using T4 DNA ligase (NEB). Products of each sequential ligation were separated by agarose gel electrophoresis, and bands of appropriate length were recovered using a commercial extraction kit (QIAGEN).

The primary cassette $L(ABC)_1R \equiv K_1$ was then ligated into a high copy-number plasmid (pYIC, Addgene plasmid 18673) using unique AflII and NdeI (NEB) restriction sites.

This generated the primary plasmid, pK_1, containing the cassette K_1 (one repeat of the ABC sequence).

Plasmids pK_1 were transformed into the DH5α strain of *E. coli* (Invitrogen) according to the manufacturer's protocol and were selected at 37°C on LB-agar plates supplemented with 25 μg/mL kanamycin. Following cell growth in liquid LB-kanamycin medium, plasmids were purified using a miniprep kit (QIAGEN). Correct incorporation of cassettes was confirmed by restriction assays (Figure 4) and sequencing (MWG Operon, Huntsville, AL, USA).

4.2.2. Doubling of Cassettes. The doubling of the cassette is illustrated schematically in Figure 4. To double the length of the (ABC) repeating unit, the pK_1 plasmid was double-digested with either BbsI and NdeI (donor) or NdeI and BsaI (recipient) (all enzymes: NEB). Appropriately sized fragments were gel purified (QIAGEN). Donor and recipient were ligated together, and the new construct (pK_2 plasmid) was transformed into DH5α cells as described above. This process was repeated to generate pK_4, pK_8, pK_{16}, and pK_{32} plasmids.

4.3. Track Elongation and Biotinylation. pK_8 plasmid (4179 bp) was digested with EagI and RsrII (NEB) to generate two fragments, of which the 3573 bp was gel purified. Lambda DNA (NEB) was digested with PspOMI to generate a 38416 bp fragment which was isolated by agarose gel electrophoresis and extraction (agarose gel-digesting preparation, GELase). This was subsequently ligated to the purified 3573 bp fragment of the pK_8 plasmid using the compatible cohesive ends from the PspOMI and EagI digest.

A biotin tag was introduced into the construct by using the 5'-GAC overhang, produced by initial digestion of the pK_8 plasmid with RsrII, and ligation to a short sticky-ended duplex DNA made from oligonucleotide 5'-biotin-CGGCATCAGAGCAGATGAC-3' and its complement 5'-ATCTGCTCTGATGCCG-3'. To probe for the presence of the biotin tag on the construct, the sample was run on a 0.7% TBE agarose gel to separate the construct from any unligated short oligonucleotides, then transferred to a nitrocellulose membrane (Trans-Blot, Bio-Rad). The membrane was washed and blocked (with 2% BSA in 1X DPBST + 0.1% Tween) and probed according to the manufacturer's protocol with horseradish peroxidase-conjugated streptavidin (Figure 6).

4.4. Labelling of a DNA Track with Fiducial Markers

4.4.1. Preparation of 300 Base-Pair Alkyne DNA. A 300 bp DNA product which incorporates alkyne-modified deoxynucleotides was prepared following a literature protocol [38]. Briefly, PCR amplification of pUC19 plasmid DNA was performed using the following primer pairs: 5'-ATGCTTCGGCCGTATGCGGTGTGAAA-3' and 5'-TCTTATGGGCCCTGTGGAATTGTGAG-3'. These primers generate EagI and PspOMI restriction sites at the two ends of the PCR product. In the PCR reaction mixture, the natural dTTP was replaced with the alkyne-modified deoxynucleotide analog, C8-alkyne-dUTP (Jena Bioscience). DNA was amplified using Pwo polymerase (Roche). Products of PCR reactions were purified using the QiaQuick PCR purification kit (QIAGEN), and their size was confirmed by agarose gel electrophoresis. Alkyne-labelled DNA ran slightly slower on a 1% agarose gel relative to control samples (Figure 8).

4.4.2. Preparation of Carboxyrhodamine 110-Labelled DNA. To 1 pmol of alkyne-labelled, 300-base pair DNA was added 10 pmol of azide-fluor 488 (carboxyrhodamine 110-PEG3-azide, Jena Bioscience). Quantitative incorporation of the fluorophore would result in one fluorophore per 30 bp of sequence. The reaction was catalyzed by the addition of copper (II) sulphate (0.5 mM) and ascorbic acid (0.5 mM) in the presence of 0.5 mM tris[(1-benzyl-1H-1,2,3-triazol-4-yl)methyl] amine (TBTA) in 50% DMSO according to a manufacturer's protocol (Lumiprobe). TBTA was added to protect the DNA from Cu(I)-mediated DNA strand breaks [39]. The reaction was incubated at room temperature for 1 hour with rotary mixing. A control reaction, in which the DNA contained all natural bases and no alkynes, was similarly treated. The reaction mixture was purified by ethanol precipitation using glycogen to aid precipitation. Appropriate labelling of the alkyne DNA was confirmed by running the sample on a 1% agarose gel and scanning the gel at 532 nm and monitoring emission at 555 nm (555/20 nm band-pass filter, Figure 7).

4.4.3. Ligation of Alkyne-Labelled DNA. Alkyne-labelled 371 bp DNA was generated by PCR using the following primers which are modified from those described above in order to further exclude alkyne-labelled nucleotides from the vicinity of the EagI and PspOMI recognition sites: 5'-TTGCTTCGGCCGTTGTACTGAGAGTGCACC-3' and 5'-TCTTGTGGGCCCTGTGGAATTGTGAGCG-3'. Alkyne-labelled DNA was digested using either EagI or PspOMI. After product clean-up (PCR purification kit, QIAGEN), a self-ligation reaction of each batch was run using T4 DNA ligase. Ligation products were run on a 1% agarose gel to confirm ligation and were compared with the self-ligations of EagI- and PspOMI-digested DNA from control 371 bp fragments lacking the alkyne modification (Figure 8).

Authors' Contribution

S. Kovacic and L. Samii contributed equally to this paper and they shared the first authorship.

Acknowledgments

References

[1] A. S. Khalil and J. J. Collins, "Synthetic biology: applications come of age," *Nature Reviews Genetics*, vol. 11, no. 5, pp. 367–379, 2010.

[2] C. Bustamante, J. F. Marko, E. D. Siggia, and S. Smith, "Entropic elasticity of λ-phage DNA," *Science*, vol. 265, no. 5178, pp. 1599–1600, 1994.

[3] P. W. K. Rothemund, "Folding DNA to create nanoscale shapes and patterns," *Nature*, vol. 440, no. 7082, pp. 297–302, 2006.

[4] A. M. Hung, C. M. Micheel, L. D. Bozano, L. W. Osterbur, G. M. Wallraff, and J. N. Cha, "Large-area spatially ordered arrays of gold nanoparticles directed by lithographically confined DNA origami," *Nature Nanotechnology*, vol. 5, no. 2, pp. 121–126, 2010.

[5] N. V. Voigt, T. Tørring, A. Rotaru et al., "Single-molecule chemical reactions on DNA origami," *Nature Nanotechnology*, vol. 5, no. 3, pp. 200–203, 2010.

[6] R. Pei, S. K. Taylor, D. Stefanovic, S. Rudchenko, T. E. Mitchell, and M. N. Stojanovic, "Behavior of polycatalytic assemblies in a substrate-displaying matrix," *Journal of the American Chemical Society*, vol. 128, no. 39, pp. 12693–12699, 2006.

[7] K. Lund, A. J. Manzo, N. Dabby et al., "Molecular robots guided by prescriptive landscapes," *Nature*, vol. 465, no. 7295, pp. 206–210, 2010.

[8] L. Samii, H. Linke, M. J. Zuckermann, and N. R. Forde, "Biased motion and molecular motor properties of bipedal spiders," *Physical Review E*, vol. 81, no. 2, Article ID 021106, 2010.

[9] L. Samii, G. A. Blab, E. H. C. Bromley et al., "Time-dependent motor properties of multipedal molecular spiders," *Physical Review E*, vol. 84, no. 3, Article ID 031111, 2011.

[10] A. K. H. Cheng, D. Sen, and H. Z. Yu, "Design and testing of aptamer-based electrochemical biosensors for proteins and small molecules," *Bioelectrochemistry*, vol. 77, no. 1, pp. 1–12, 2009.

[11] D. Rueda and N. G. Walter, "Fluorescent energy transfer readout of an aptazyme-based biosensor," *Methods in Molecular Biology*, vol. 335, pp. 289–310, 2006.

[12] T. C. Marsh and E. Henderson, "G-wires: self-assembly of a telomeric oligonucleotide, d(GGGGTTGGGG), into large superstructures," *Biochemistry*, vol. 33, no. 35, pp. 10718–10724, 1994.

[13] C. Lin, Y. Liu, S. Rinker, and H. Yan, "DNA tile based self-assembly: building complex nanoarchitectures," *ChemPhysChem*, vol. 7, no. 8, pp. 1641–1647, 2006.

[14] E. H. C. Bromley, N. J. Kuwada, M. J. Zuckermann et al., "The tumbleweed: towards a synthetic protein motor," *HFSP Journal*, vol. 3, no. 3, pp. 204–212, 2009.

[15] N. J. Kuwada, G. A. Blab, and H. Linke, "A classical Master equation approach to modeling an artificial protein motor," *Chemical Physics*, vol. 375, no. 2-3, pp. 479–485, 2010.

[16] A. Cooper, A. McAlpine, and P. G. Stockley, "Calorimetric studies of the energetics of protein-DNA interactions in the E. coli methionine repressor (MetJ) system," *FEBS Letters*, vol. 348, no. 1, pp. 41–45, 1994.

[17] I. D. Parsons, B. Persson, A. Mekhalfia, G. M. Blackburn, and P. G. Stockley, "Probing the molecular mechanism of action of co-repressor in the E.coli methionine repressor-operator complex using surface plasmon resonance (SPR)," *Nucleic Acids Research*, vol. 23, no. 2, pp. 211–216, 1995.

[18] R. W. Schevitz, Z. Otwinowski, A. Joachimiak, C. L. Lawson, and P. B. Sigler, "The three-dimensional structure of trp repressor," *Nature*, vol. 317, no. 6040, pp. 782–786, 1985.

[19] R. G. Zhang, A. Joachimiak, C. L. Lawson, R. W. Schevitz, Z. Otwinowski, and P. B. Sigler, "The crystal structure of trp aporepressor at 1.8 Å shows how binding tryptophan enhances DNA affinity," *Nature*, vol. 327, no. 6123, pp. 591–597, 1987.

[20] A. Nagadoi, S. Morikawa, H. Nakamura et al., "Structural comparison of the free and DNA-bound forms of the purine repressor DNA-binding domain," *Structure*, vol. 3, no. 11, pp. 1217–1224, 1995.

[21] M. A. Schumacher, K. Y. Choi, H. Zalkin, and R. G. Brennan, "Crystal structure of LacI member, PurR, bound to DNA: minor groove binding by α helices," *Science*, vol. 266, no. 5186, pp. 763–770, 1994.

[22] J. Carey, "Gel retardation at low pH resolves trp repressor-DNA complexes for quantitative study," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 85, no. 4, pp. 975–979, 1988.

[23] K. L. B. Borden, P. Beckmann, and A. N. Lane, "Determination of the orientations of tryptophan analogues bound to the trp repressor and the relationship to activation," *European Journal of Biochemistry*, vol. 202, no. 2, pp. 459–470, 1991.

[24] K. Phillips and S. E. V. Phillips, "Electrostatic activation of Escherichia coli methionine repressor," *Structure*, vol. 2, no. 4, pp. 309–316, 1994.

[25] T. Fazio, M. L. Visnapuu, S. Wind, and E. C. Greene, "DNA curtains and nanoscale curtain rods: high-throughput tools for single molecule imaging," *Langmuir*, vol. 24, no. 18, pp. 10524–10531, 2008.

[26] J. Gorman, T. Fazio, F. Wang, S. Wind, and E. C. Greene, "Nanofabricated racks of aligned and anchored DNA substrates for single-molecule imaging," *Langmuir*, vol. 26, no. 2, pp. 1372–1379, 2010.

[27] S. Forrest, "Genetic algorithms: principles of natural selection applied to computation," *Science*, vol. 261, no. 5123, pp. 872–878, 1993.

[28] P. Willett, "Genetic algorithms in molecular recognition and design," *Trends in Biotechnology*, vol. 13, no. 12, pp. 516–521, 1995.

[29] M. Affenzeller, S. Winkler, S. Wagner, and A. Beham, *Genetic Algorithms and Genetic Programming: Modern Concepts and Practical Applications*, Chapman & Hall/CRC, New York, NY, USA, 2009.

[30] J. W. Eaton, *GNU Octave Manual*, Network Theory Limited, 2002.

[31] N. R. Markham and M. Zuker, "DINAMelt web server for nucleic acid melting prediction," *Nucleic Acids Research*, vol. 33, no. 2, pp. W577–W581, 2005.

[32] N. R. Markham and M. Zuker, "UNAFold: software for nucleic acid folding and hybridization," *Methods in Molecular Biology*, vol. 453, pp. 3–31, 2008.

[33] A. M. Albertini, M. Hofer, M. P. Calos, and J. H. Miller, "On the formation of spontaneous deletions: the importance of short sequence homologies in the generation of large deletions," *Cell*, vol. 29, no. 2, pp. 319–328, 1982.

[34] J. W. Eaton, D. Bateman, and S. Hauberg, *GNU Octave Manual Version 3*, Network Theory Limited, 2008.

[35] C. Liébecq, *Compendium of Biochemical Nomenclature and Related Documents*, Portland Press, 1992.

[36] D. Kozen, "A completeness theorem for Kleene algebras and the algebra of regular events," in *Proceedings of the 6th Annual IEEE Symposium on Logic in Computer Science (LICS '91)*, pp. 214–225, July 1991.

[37] T. Stubblebine, *Regular Expression Pocket Reference*, O'Reilly Media, Inc., 2003.

[38] J. Gierlich, K. Gutsmiedl, P. M. E. Gramlich, A. Schmidt, G. A. Burley, and T. Carell, "Synthesis of highly modified DNA by a combination of PCR with alkyne-bearing triphosphates and click chemistry," *Chemistry*, vol. 13, no. 34, pp. 9486–9494, 2007.

[39] T. R. Chan, R. Hilgraf, K. B. Sharpless, and V. V. Fokin, "Polytriazoles as copper(I)-stabilizing ligands in catalysis," *Organic Letters*, vol. 6, no. 17, pp. 2853–2855, 2004.