

Molecular Shift Register and its Utilization as an Autonomous DNA Synthesizer

Ilya Baskin,^{1,2,4} Stav Zaitsev,^{1,2,4} Doron Lipson,^{1,3} Rachel Gilad,¹ Kinneret Keren,^{1,4}
Gidi Ben-Yoseph,^{1,4} and Uri Sivan^{1,4,*}

¹Department of Physics, Technion-Israel Institute of Technology, Haifa 32000, Israel

²Department of Electrical Engineering, Technion-Israel Institute of Technology, Haifa 32000, Israel

³Department of Computer Science, Technion-Israel Institute of Technology, Haifa 32000, Israel

⁴The Russell Berrie Nanotechnology Institute, Technion-Israel Institute of Technology, Haifa 32000, Israel

(Received 1 July 2006; published 16 November 2006)

A novel algorithmic approach to the synthesis of fairly long DNA molecules with nonrecurring sequences is demonstrated. The scheme exploits chemical embodiment of shift registers (SR) to execute algorithms similar to those used to generate pseudorandom numbers on a computer. Single stranded DNA molecules guide the synthesis of double stranded DNA according to the SR truth table. The SR logic facilitates an exponentially smaller synthesis effort compared with all other strategies. A redundancy based scheme, similar to those used in communication, is utilized to suppress synthesis errors.

DOI: 10.1103/PhysRevLett.97.208103

PACS numbers: 87.14.Gg

The synthesis of long DNA molecules remains a challenge in biotechnology. One of the successful strategies, assembly polymerase chain reaction (PCR) [1], utilizes synthetic oligos, which together span the full length of the desired molecule. Complete genes and plasmids were assembled this way. Programmed synthesis is also inherent to DNA computing. Adleman's pioneering work [2] showed how oligonucleotides representing segments of a graph could be ligated to yield a long molecule corresponding to a path on that graph. The same ligase assembly was later shown to correspond to finite state machines [3] and regular languages [4]. The same is true for assembly PCR [5]. In all these instances, a long DNA molecule was assembled from its pieces based on the recognition of a given piece by the subsequent one. Following any of these strategies, the assembly of an N base long molecule with distinct p -long segments requires $O(N/p)$ oligos. These approaches, therefore, quickly become impractical when a rich variety of distinct molecules or addresses along a given molecule are needed for, e.g., the construction of an elaborate template for molecular electronics [6–9]. Here we introduce an exponentially more economic synthesis strategy based on chemical realization of molecular shift registers.

An autonomous binary p -shift register (p -SR) is a computing machine with 2^p internal states represented by a linear array of p cells, each occupying 1 bit, x_i ($i = 1, \dots, p$). In each step a binary function, $f(x_1, x_2, \dots, x_p)$, is computed and its value is inserted into cell p . Simultaneously, x_j is shifted to cell $j - 1$; ($j = 2, \dots, p$). On printing x_1 to a tape, a long periodic binary sequence is generated. Electronic SRs are utilized in many applications including secure communication, small signal recovery, and sequence generation [10]. Here we show that molecular SRs can be realized and utilized for autonomous synthesis of DNA molecules whose sequence is uniquely determined by a chemical embodiment of the function $f(x_1, x_2, \dots, x_p)$.

Consider a 3-SR with $x_{n+1} = f(x_{n-2}, x_{n-1}, x_n) = x_{n-2} \oplus x_n$ ($\oplus \equiv \text{XOR}$) and an initial setting (seed) $x_1, x_2, x_3 = 001$. Repetitive application of f generates the sequence 001110100111010... The sequence is periodic with a period 7 and any of the $L \geq 3$ bit long consecutive strings in a period is different from the rest. Generally it is known [10] that for any p , one can find a SR with a linear feedback function [11], $f = \sum_{i=1}^p \alpha_i x_i$; $\alpha_i \in \{0, 1\}$ (the sum is mod 2), that generates a sequence of maximal period $2^p - 1$ bits [12] with no repetition of strings of lengths $L \geq p$ within a period.

We now show how to implement such an autonomous molecular SR using DNA. Imagine a DNA molecule whose Watson-Crick rules are that 1 binds exclusively to its complementary bit, $\bar{1}$, but not to 1, 0, or $\bar{0}$. Similarly, 0 binds to $\bar{0}$ but not to 0, $\bar{1}$, or 1. We translate the function $f(x_1, x_2, x_3) = x_1 \oplus x_3$ to an equivalent truth table (left three columns in Table I) and embody it by the mixture of the seven [12] possible 4-bit rule strands, $(\bar{x}_1, \bar{x}_2, \bar{x}_3, (\bar{x}_1 \oplus \bar{x}_3))$, listed in the right column of Table I.

The SR sequence is generated by thermally cycling a mixture containing the 7 rule strands, a “seed” strand, and polymerase. For simplicity, assume the rule strands are synthesized with ddDNA at their 3' end and are therefore not elongated in the process. In the first annealing step, some of the first, $\bar{0}\bar{0}\bar{1}\bar{1}$, rule strands bind to seed molecules, leaving an $\bar{1}$ overhang [Fig. 1(i) and 1(ii)] which is

TABLE I. Truth table and rule strands corresponding to the first example.

x_1	x_3	f	Rule Strand
0	0	0	$\bar{0}\bar{1}\bar{0}\bar{0}$
0	1	1	$\bar{0}\bar{0}\bar{1}\bar{1}, \bar{0}\bar{1}\bar{1}\bar{1}$
1	0	1	$\bar{1}\bar{0}\bar{0}\bar{1}, \bar{1}\bar{1}\bar{0}\bar{1}$
1	1	0	$\bar{1}\bar{0}\bar{1}\bar{0}, \bar{1}\bar{1}\bar{1}\bar{0}$

$$5' \text{GCATGCGCCCGTCAGGCG}00111(0100111)_n 01001\text{CTGCAG} \text{ with } n = 0, 1, \dots \quad (1)$$

seed primer
↙
↘ complementary to stop primer

The elongation products, diluted 1000-fold to reduce the concentration of rule strands, are PCR amplified with two primers, identical to the first 19 nucleotides of the seed and to the last 19 nucleotides of the stop primer.

The resulting PCR products, run against a standard ruler in a polyacrylamid gel, are depicted in Fig. 2(a). Four bands corresponding to formula (1) with $n = 0, 1, 2, 3$ are clearly resolved. DNA is extracted from each band and sequenced with a primer identical to the first 19 nucleotides on the 5' end of the seed primer. The sequences prove the bands identification with the respective n values in (1).

As demonstrated by Fig. 2(b), after 100 elongation cycles we were able to resolve 10 bands. Our automaton thus synthesizes at least 204 bases at a remarkable fidelity. The $n = 9$ sequence comprises 10 periods, 21 bases each, with exactly one repetition of each 3-bit (or longer) address per period. Direct sequencing of the bands confirmed our results up to $n = 6$. The small material quantities in higher bands were insufficient for reliable sequencing. An example for a sequencing trace [the 138 base-pair band of Fig. 2(b)] is depicted in Fig. S1 in Ref. [17]. The high fidelity of our automaton is reflected in perfect matching of the sequences with formula (1), the low background in sequencing, and the absence of unexpected bands.

To rule out the possibility that SR molecules were synthesized in paths other than the planned one we have run in parallel to all processes identical reactions depleted of seed molecules. No SR sequences were ever observed in those control tubes. The possibility that SR sequences were generated in the PCR amplification step was ruled out by the fact that the lengths of the produced SR sequences were monotonic in the number of elongation cycles and independent of the number of PCR amplification cycles. Comparison between the product length of 45 and 100 elongation cycles is provided by Figs. 2(a) and 2(b), respectively.

The SR efficiency was estimated by running the elongation products in real time PCR against a calibrated standard. We found this way that more than 20% of the seed molecules matured to the point were a stop primer terminated their elongation.

We find that elongation can be carried out at a constant temperature, $T \approx 72^\circ\text{C}$, without thermal cycling [22]. This thermal ratchet mode of operation is possible since, in contrast to conventional PCR, the tape is elongated rather than the primer (the latter is elongated by at most 2 bases). Consequently, the rule strand-tape melting temperature remains low and the primers may detach and attach spontaneously. Whenever the right rule strand binds to the tape the latter is elongated by 1 bit. Errors bind more rarely and for shorter times. Binding without overhang merely slows synthesis by occupying the tape molecule. The next examples were realized at a constant temperature.

A maximal 4-SR sequence using 15 7-bit rule strands [23] (6 bit rules plus 1 function bit) is demonstrated next. The feedback function used was $x_{n+1} = x_n \oplus x_{n-3}$. The $2^4 - 1 = 15$ period sequence reads $0011110(101100100011110)_n 1011$; $n = 0, 1, \dots$ so that each 4-bit address (or longer) appears exactly once per period. Remarkably, the truth table dimension, and hence the minimal synthesis effort, remains 2^2 though the number of distinct addresses is $2^4 - 1$. As seen in Fig. 2(c), five bands corresponding to $n = 0-4$ are clearly resolved. Direct sequencing confirmed the bands identification.

The binary realizations presented thus far were extremely conservative with respect to the maximal information that can be encoded in a SR sequence of a given length. With three nucleotides per bit it is possible, in

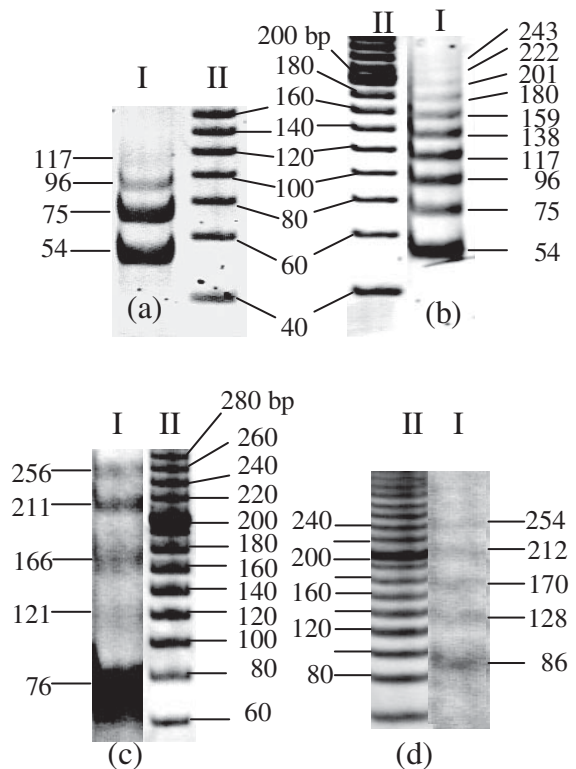


FIG. 2. (a) Lane I—product after 45 elongation cycles, 5 cycles with stop primer, and PCR amplification. The four bands correspond to (1) with $n = 0, 1, 2, 3$, namely, 54, 75, 96, and 117 base long sequences. Lane II ruler. (b) Lane I—same as (a) with 100 elongation cycles followed by PCR and filtering out short sequences (Microcon YM-10, Millipore Corporation, Bedford, MA, USA). Ten bands corresponding to (1) with $n = 0, 1, 2, 3, 4, 5, 6, 7, 8, 9$ are resolved. Lane II ruler. (c) Four-shift register realized with 7-bit rule strands. 2h reaction time at a constant temperature of 72°C . Lane I—shift register product. The five resolved bands are indicated. Lane II ruler. (d) Same as (c) for partial three-shift register with four-letter alphabet. The period comprises 14 letters (42 bases).

principle, to code an alphabet with $4^3 = 64$ letters and, hence, a sequence with a period $64^p - 1$. Practically, such a maximal coding is likely to suffer from a high error rate due to interference between the letters, dimerization, etc. The identification of an optimal alphabet for our purposes presents an outstanding theoretical and experimental challenge. Here we prove that SRs with 4-letter alphabet can be realized. The function $x_{n+1} = \text{mod}_4[x_n + x_{n-2}]$ was realized with 14 rules [24] that generate the nonmaximal sequence $03110(12231323203110)_n1223132320$ $n = 0, 1, \dots$. Five bands corresponding to $n = 0, 1, 2, 3, 4$ are clearly resolved in the gel displayed in Fig. 2(d). The first three bands were confirmed by sequencing.

To generate a large library of different sequences one synthesizes once all possible rules of a given length. The products can later be used to synthesize any desired SR sequence by mixing subsets of rules. For p -SRs with k -letter alphabet the full library (linear and nonlinear rules) comprises k^{p+1} strands while the number of maximal sequences of length k^p equals [25] $[(k-1)!]^{k^{p-1}} k^{k^{p-1}-p}$. For instance, the full libraries of 4 and 5-SR rules with $k = 3$ comprise 243 and 729 strands, respectively. Their various combinations yield, respectively, more than 10^{20} and 10^{60} distinct maximal SR sequences. Thus, for the synthesis of $s = k^{p+1}$ sequences it is possible to generate an exponentially large number of sequences, $(k!)^{s/k^2} k s^{-1}$.

SRs may hence be used to generate either a library of sequences or a long molecule at an exponentially small synthesis effort. If one considers truly complex DNA based assembly with a large number of different components the proposed approach might turn essential. Other applications of electronic SRs in communication and signal processing may also be adapted to their chemical embodiment. For instance, automatic translation of communication messages to new ones with better error immunity may find applications in hybridization experiments.

Research was supported by the Israeli Science Foundation.

*To whom correspondence should be addressed.

Electronic address: phsivan@tx.technion.ac.il

- [1] W.P. Stemmer *et al.*, *Gene* **164**, 49 (1995).
 [2] L.M. Adleman, *Science* **266**, 1021 (1994).
 [3] D. Boneh *et al.*, *Discrete Appl. Math.* **71**, 79 (1996).
 [4] E. Winfree, X. Yang, and N.C. Seeman, in *DNA Based Computers II*, edited by Laura Landweber and Eric Baum, DIMACS Series in Discrete Mathematics and Theoretical Computer Science Vol. 44 (DIMACS, Piscataway, NJ, 1996), p. 191.
 [5] E. Winfree, Caltech Computer Science Technical Reports No. 1998.23, 1998.
 [6] J.J. Storhoff and C.A. Mirkin, *Chem. Rev.* **99**, 1849 (1999).
 [7] P.W.K. Rothmund, *Nature (London)* **440**, 297 (2006).
 [8] E. Braun, Y. Eichen, U. Sivan, and G. Ben-Yoseph, *Nature (London)* **391**, 775 (1998).
 [9] K. Keren *et al.*, *Science* **297**, 72 (2002).
 [10] Solomon W. Golomb, *Shift Register Sequences* (Aegean Park, Laguna Hills, CA, 1982).
 [11] Although we realize linear shift registers here, the automaton should work equally well with nonlinear feedback functions.
 [12] The zero string should be avoided since it maps onto itself by any linear feedback function.
 [13] M. Hagiya *et al.*, *DNA Based Computers III*, DIMACS Series in Discrete Mathematics and Theoretical Computer Science Vol. 48 (DIMACS, Piscataway, NJ, 1997), p. 57.
 [14] K. Sakamoto *et al.*, *Science* **288**, 1223 (2000).
 [15] J. Khodor and D.K. Gifford, *Theory Comput. Systems* **35**, 483 (2002).
 [16] D. Faulhammer, A.R. Cukras, R.J. Lipton, and L.F. Landweber, *Proc. Natl. Acad. Sci. U.S.A.* **97**, 1385 (2000).
 [17] See EPAPS Document No. E-PRLTAO-97-007647 for a detailed discussion of the error suppression mechanism, protocols of the various procedures used in the experiment, and a sample sequencing trace. For more information on EPAPS, see <http://www.aip.org/pubservs/epaps.html>.
 [18] A three-shift register realized with 4-bit rule strands might have worked as well but we chose to suppress errors by using rule strands longer than the minimal ones.
 [19] A 5-bit rule is constructed by adding the two preceding bits of the sequence to the 3' end of the corresponding 3-bit rule.
 [20] N. Sugimoto, S. Nakano, M. Yoneyama, and K. Honda, *Nucleic Acids Res.* **24**, 4501 (1996).
 [21] For brevity we use notation that mixes bases with bits. Each bit represents three bases.
 [22] K. Sakamoto *et al.*, *BioSystems* **52**, 81 (1999).
 [23] Fifteen 21-mer rule strands 3'0011110, 3'0111101, 3'1111010, 3'1110101, 3'1101011, 3'1010110, 3'0101100, 3'1011001, 3'0110010, 3'1100100, 3'1001000, 3'0010001, 3'0100011, 3'1000111, 3'0001111. A 36-mer seed strand 5'GCATGCGCCCGTCAGGCG001111. A 37-mer stop primer 3'101100GCGCCAGGACGCGGACGTC, and two PCR primers 5'GCATGCGCCCGTCAGGCG, 3'GCGCCAGCACGACGCGGACGTC.
 [24] Fourteen 18-mer rule strands 3'031101, 3'311012, 3'110122, 3'101223, 3'012231, 3'122313, 3'223132, 3'231323, 3'313232, 3'132320, 3'323203, 3'232031, 3'320311, 3'203110. "0" = 5'TGG, "1" = 5'GTC, "2" = 5'GCT, "3" = 5'CCT. A 37-mer seed strand 5'-CTGCAGGGACCAGGTACTGCGT03110, a 37-mer stop primer 5'CGTACGGGCCAGACGACCG023231, and two PCR primers 5'-CTGCAGGGACCAGGTACTGCG, 5'CGTACGGGCCAGACGACCG.
 [25] H. Fredricksen, *SIAM Rev.* **24**, 195 (1982).