2.4 DNA Computing on Surfaces: the Chemical Implementation


I. Introduction:

In this chapter, we explore the mechanics behind a DNA computer, integrating biomolecular approaches into the logical world of computation. We focus on setting up the various parts of the DNA computer; the actual operations are more fully explored in chapter 3.5.

The premise of using DNA as a computational tool is based on its massively parallel-search ability, which, in principle, makes possible the solution of cumbersome problems, such as Hamiltonian path, \(^1\) non-polynomial (NP), \(^2\) and maximum clique problems. \(^3\) Although building the device may appear to be less complex than integrated-circuit fabrication, the computation itself involves numerous potentially unwieldy laboratory steps. In an effort to minimize these unpleasant realities of dealing with a biologically based computation system, we have developed a surface-based technique. \(^4\) A surface-based approach may appear inadvisable because of the loss of information density in a 2-dimensional system compared to a 3-dimensional one. One milliliter of a one-millimolar solution of DNA contains \(10^{17}\) molecules of DNA, but a 10 x 10 centimeter square surface, in which a DNA molecule may occupy 5 x 5 nanometers square on the surface, holds only about \(10^{14}\) molecules of DNA. \(^4\) This limitation is countered by the greater simplicity and reliability of the surface-based approach, and also the easy scalability to larger sample areas and multiple surfaces.
We favor the solid-support approach over a solution-based DNA computation because of (1) the ease of sample handling, (2) the decrease of losses during handling, (3) the reduction of interference between oligonucleotides, (4) facile purification, and (5) potential to integrate DNA "computer surfaces" and read-out arrays into silicon-based processing. (1) With DNA bound to a surface, the computational operations of the computer become very simple. A flat surface with the DNA tethered to it is introduced to solutions, for enzymatic ligations or for hybridizations, and later it is washed to remove solutions. (2) When the DNA is attached to the surface, the loss of DNA between one step and another is minimal. In a solution-based DNA computational system, every transfer from one test tube to another leaves residual liquid behind, and in that liquid, some fraction of the DNA mixture, thus introducing error to the system. (3) Once immobilized on a surface, the oligonucleotides cannot interact with each other, whereas in solution, two complementary single-stranded sequences can form duplexes. (4) Purification of DNA sequences, necessary at each step, is as simple as washing the surface with water to remove unwanted buffers, salts, or other reaction products. The ease of going from one reaction to the next contrasts with the extraction and purification of the DNA mixture from the reaction solution in solution-based computation. Separation in a solution-based system is usually achieved through gel electrophoresis, and dialysis is necessary sometimes to remove unwanted salts; both procedures are tedious and time consuming. (5) DNA computation on a surface can benefit from the wealth of tools available in silicon processing, designed for integrated-circuit fabrication, including patterning, etching, and transporting of silicon wafers, eliminating the need to design new materials-processing technology.
To have confidence in the data and to achieve accurate, consistent results, one must use high-purity DNA. In addition, the conditions for each enzymatic ligation or hybridization are specific and it is essential to be able to control the environment (i.e., the buffer solutions) of these reactions. A surface-based approach allows for a highly controllable system and keeps intrinsic errors to a minimum.

II. Synthesis and Preparation:

In the present implementation of surface-based DNA computation, it is necessary to immobilize the DNA sequences on the solid support, select or MARK a specific subset through hybridization, remove or DESTROY the unwanted sequences, then deselect or UNMARK the specific subset, and finally analyze or READOUT the sequences that remain on the surface. These operations are illustrated in Figure 1.  

⁵
Figure 1: A schematic outline of the DNA operations for surface-based DNA computation. Generation involves creating a large pool of different DNA sequences to represent the different possible solutions to a problem. We immobilize the strands to a surface, MARK a specific subset of sequences with its complementary sequence, and then DESTROY the remaining single-stranded sequences on the surface. Afterwards, the complementary sequences are UNMARKED. Steps 3, 4
Design and Generation:

In previous work, it has been demonstrated that the conditions employed for single-nucleotide encoding of information result in low hybridization efficiency. Accordingly, instead of encoding a bit (0 or 1) of information at a single nucleotide, a “word” design strategy has been developed to store 4 to 8 bits of information in an 8-nucleotide region (a “word”) in which each “word” differs from every other word by at least 4 base pairs. The DNA words have the following design:

\[
5'\text{-FFFFVVVVVVFFFF-3'},
\]

where \(F\) is a fixed base and \(V\) is a variable base. There are \(4^8\) (64,536) possible 8mers to encode as words, but the DNA words are further limited by their base content. The ratio of the guanine and cytosine bases (G/C content) to the adenine and thymine bases significantly affects the temperature at which the duplex of a given sequence dissociates into two separate strands (melting temperature, \(T_m\)). A G/C content of 50% was chosen to provide uniform thermodynamic stability of the duplex. A minimum sequence difference of four bases out of eight for each word was chosen to ensure a large difference in \(T_m\) between a perfectly matched duplex (~70°C) and a mismatched duplex (25-30°C). Using these conditions as a prerequisite and rejecting palindromic and slide sequences, we have identified a set of 108 8mers with a G/C content of 50%. For further discussion, see chapter 3.5.

From the set of 108 8mers, we approach the problem of generating a combinatorial mixture of DNA sequences using a split-pool strategy. After synthesizing
four DNA words in four separate test tubes on support particles via controlled-pore-glass (CPG) synthesis, a standard technique in oligonucleotide synthesis, these test tubes are opened, mixed, and then re-distributed into four separate test tubes to add another word on the strands, a different word in each test tube. After two rounds of mixing and redistribution of the DNA strands, a set of all 16 different possible combinations is achieved, as seen in Figure 2. If such a strategy were repeated for 108 words for 6 cycles, $1.6 \times 10^{12}$ ($10^{86}$) different strands of 8mers would be generated. After six cycles, six DNA words on a single strand, each word 16 base pairs long, would yield a 96mer. The CPG synthesis has an efficiency rate of 0.995 for each base addition, and so a strand of 96 bases long will have a 62% yield of the correct sequence. Clearly long DNA strands become less pure; a ~100-base long sequence is near the limit.
Figure 2: Generation of the DNA strands. (a) Four DNA words (encoded with the different patterns) are synthesized separately. (b) The DNA words are mixed together and (c) redistributed into four new test tubes for the attachment of a second word onto the DNA strands. (d) The set of DNA strands generated is represented with the pattern scheme. There are 16 different DNA sequences after two rounds with four words each.

Immobilization:

There are many methods to tether DNA strands to a surface. The biotechnology industry has a flourishing group of DNA-array and micro-array makers. Applications of arrays lie in high-throughput analyses for, to name a few, gene expression, gene variation, toxicology, and forensic studies. We use self-assembled monolayers (SAMs) on a gold surface for the controlled attachment of DNA to a surface.

Alkanethiol monolayers, a class of SAMs that react with gold and other noble-metal surfaces, self-assemble to form robust, highly ordered films. The use of ω-functionalized alkanethiols, whether commercially available or specifically synthesized, allows the surface reactivity to be tailored via the choice of terminal group. The DNA attachment takes advantage of both the ordered-monolayer characteristic of alkanethiols and the selective reactivity of the terminal group.

In Figure 3, the process of attaching DNA to a surface with alkanethiols is illustrated schematically. A glass slide or silicon wafer with a 1000Å-thick evaporated gold film deposited on an adhesion layer of 50Å of chromium is immersed in 1mM ethanolic mercapto-undecanoic acid (MUA) solution for at least 18 hours to form a self-assembled monolayer on the surface. The sample is then covered with 1mg/mL poly-L-lysine (PL) pH 8.0 for 30 minutes to form an electrostatic bond between the carboxylic acid groups of the MUA molecules and the amine groups of the polylysine. After reacting the surface with 1mM sulfosuccinimidyl 4-(N-maleimidomethyl) cyclohexane-1-
carboxylate (SSMCC) for 30 minutes, the 5’-thiol-modified DNA is placed onto the slide and allowed to react for at least 6 hours in a humid chamber at room temperature. The surface is rinsed with water and soaked for 1 hour in a solution of 300mM NaCl, 20mM sodium phosphate, 2mM EDTA, and 6.9mM sodium dodecyl sulfate (2X SSPE/0.2% SDS) at which time the sample is ready for the first MARK operation, or hybridization of a specific subset of DNA molecules.\textsuperscript{5}

Figure 3: DNA attachment to the surface via alkanethiols.
MARK:

As shown in Figure 4, to MARK or hybridize the DNA strands that satisfy the first clause of a satisfiability (SAT) problem represented in a DNA-computation system, a solution of complementary DNA sequences that represent the DNA words that satisfy the first clause is exposed to the combinatorial mixture of DNA sequences. The hybridization to the attached DNA-probe molecules is accomplished by exposure to a 2μM solution of target oligonucleotides in 2X SSPE/0.2% SDS buffer. A 30μl drop of a complement solution containing DNA oligonucleotides that are complements to some of the initial DNA strands is placed onto the slide and then spread over the entire surface by placing a clean cover slip on top of the sample. Hybridization adsorption is allowed to proceed for 30 minutes, after which the sample is rinsed with water and is immersed in a solution of 2X SSPE/0.2% SDS buffer without the complement for 10 minutes before any further reactions.
Figure 4: DNA Structure. The four bases are arranged along the sugar-phosphate backbone in a particular order (the DNA sequence). Adenine (A) pairs with thymine (T), while cytosine (C) pairs with guanine (G). Weak bonds between the bases hold the two DNA strands together. This pairing is called hybridization, and a particular sequence has a corresponding complementary sequence. It is this characteristic of specific hybridization that we use during MARK to find the encoded variable for a given SAT problem.
DESTROY:

The DESTROY operation consists of adding an exonuclease specific for single-stranded DNA. Every unmarked strand is destroyed, leaving on the surface the MARKED, or hybridized, DNA molecules. For the DESTROY operation, the DNA strands on the surface are exposed to 20 U of the single-strand-specific enzyme *Escherichia coli* Exonuclease I in a pH 9.5 buffer solution consisting of 67mM glycine, 6.7mM MgCl₂, 10mM 2-mercaptoethanol, 1M NaCl and 100µm/mL BSA. Enzymatic digestion proceeds for 3 hours at room temperature; afterwards, the surface is rinsed with water.
Figure 5: The systematic destruction of single-stranded DNA molecules on a surface through the DESTROY operation. W1 signifies word 1; W2, word 2; W3 word 3; W4, word 4. For the sequences of each word, see figure 6.5

UNMARK:

To remove the complement or UNMARK, the sample is placed in 8.3M urea solution at 37°C for 20 minutes and is washed thoroughly with water. The sample is then ready for a second hybridization or MARK to target sequences that might represent solutions to the second clause in the SAT problem and any other DNA computation operations. 5

Figure 6: Selective MARK and UNMARK of DNA strands on a surface. When a DNA strand is MARKed with its fluorescent complement, signal from the fluorescent complement is detected
spectroscopically. When a DNA strand remains single stranded, there is no signal detected. W1 signifies word 1; W2, word 2; W3 word 3; W4, word 4.5

READOUT:

In READOUT, we determine the identity of the DNA strands remaining on the surface. These strands represent the answer(s) to the problem to be solved by DNA computation. In the previous steps, we have employed unaddressed surfaces, in which we do not know the location of any particular sequence on the surface. The operations of DNA computation sort the DNA into sequences that represent answers to the problem to be computed, and those that do not (the latter ones have been removed in the various computation steps), as seen in Figure 7. But the process so far, although it has found the sequences that represent answer(s) to the problem, does not identify what those sequences are. To do so, readout arrays, in which DNA sequences are placed at specific locations of the array, are synthesized.14 For READOUT of the first clause of a solution, we reapply the complementary strands that satisfy the first clause, but this time the complementary sequences contain primer oligonucleotides. When the complementary strands are removed, the polymerase chain reaction (PCR) is used with fluorescent primers to make many copies, analogous to signal amplification via a photomultiplier tube (PMT) in spectroscopy. The fluorescent products are hybridized to an addressed array and the signal is recorded. These steps are repeated for each clause. Now, we have a small combinatorial set of solutions. Let’s call the smaller combinatorial set the starred set, as depicted in figure 7. Out of the starred set, we can further winnow the solutions down by use of a conventional computer. For further discussion, see chapter 3.5.
Figure 7: READOUT strategy. Let the entire circle represent the combinatorial-solution space. Let circle A represent the DNA words that satisfy the first clause of a SAT problem; circle B, the DNA words that satisfy the second clause; circle C, the third clause; circle D, the fourth clause. After the DNA computation operations are performed on the surface, only those sequences in the starred (*) region will remain, as they are the only ones that satisfy all the clauses. It is upon this subset that we perform READOUT to identify the sequences that remain from the large combinatorial set.

III. Automation/Scale-up:
While we have developed a DNA computation system and demonstrated its power to solve a three-variable satisfiability (SAT) problem, we also must keep in mind automation and scale up of the system. If DNA computation is to be a viable tool for computationally difficult problems, it must accommodate large problems, not just small demonstrations.

Generation, immobilization, MARK, UNMARK, and DESTROY already have scale-up built into their operations. In generation, as discussed previously, we already have a strategy to generate $1.6 \times 10^{12}$ different DNA strands using existing technology, and DNA sequence machines exist that synthesize in parallel. For MARK, the hybridization process for one surface covered with DNA is a highly parallel process. The de-hybridization of a surface covered with DNA, UNMARK, will not vary as the number of different strands on the surface increases. DESTROY, once again, is a highly parallel process.

For READOUT, the method to encode the information greatly affects the efficiency of the operation. For example, with our word strategy, an addressed array is required for each separate tier of words in a multi-word sequence. If a DNA strand has two separate words, then two addressed arrays are required because the fixed bases vary from one layer of words to the next, as seen in equation 2,

$$5'-F_1F_1F_1VVVVVVVF_1F_1F_1F_2F_2F_2F_2VVVVVVVF_2F_2F_2F_2-3'$$

where $F_n$ represents a fixed base of level n and V represents a variable base of the word on the tier. We can PCR amplify the words one layer at a time of a multi-word sequence
by selecting a primer specific to the fixed nucleotides of each tier, and then applying the
PCR products to the addressed array with the corresponding fixed nucleotides of the
word. Such repetition would lend itself well to automation. For other READOUT
strategies, individually sequencing each DNA answer might be necessary. As we see
with the human-genome project, large-scale sequencing of many DNA sequences is
possible.¹⁶

IV. Conclusions

We have outlined procedures for surface-based DNA computation, and identified
and explained the necessary steps. So far, we have solved small computational problems,
in particular, a 3-variable SAT problem,¹⁵ to test the methodology and to provide a proof
of principle. In the process, we have also made strides in understanding surface reactions
for the DNA attachment, enzymatic manipulations¹⁷ and PCR amplification.¹⁵ These
reactions have applications in other areas, such as genomics, single-nucleotide-
polymorphism (SNP) detection, and drug research.

Within a few short years after the introduction of the field, we have made great
strides in the realization of a practical DNA-computation system. Given that silicon-
based computers have achieved their status after a slow start and more than 50 years of
development, it is not unreasonable to predict that DNA computation will, after a period
of gestation for the next several years, become a viable technology that some day may
rival silicon computers for specific types of problems.

¹ Adleman, L.M. Molecular computation of solutions to combinatorial problems Science
266, 1021 (1994).


12 http://www.scientific.org/tutorials/articles/riley/riley.html


14 Gillmor, S.D. et al *Langmuir* Submitted for publication


16 http://www.apbiotech.com/company/company_index.html