A Surface-Based Approach to DNA Computation *

Qinghua Liu, Zhen Guo, Anne E. Condon, Robert M. Corn, Max G. Lagally, Lloyd M. Smith

University of Wisconsin
Madison, WI 57306 USA

March 11, 1996

Abstract

A new model of DNA-based computation is presented. The main difference between this model and that of Adleman is in manipulation of DNA strands that are first immobilized on a surface. This approach greatly reduces losses of DNA molecules during purification steps. A simple, surface-based model of computation is described and it is shown how to implement an exhaustive search algorithm for the SAT problem on this model.

Partial experimental progress in solving a 5-variable SAT instance is described, and possible extensions of our model that allow general computations are discussed.

*Liu, Guo, Corn and Smith are in the Chemistry Department, Condon is in the Computer Sciences Department and Lagally is in the Materials Sciences Department. Email address for further communication: condon@cs.wisc.edu.

1 Introduction

Adleman [1] and subsequently Lipton [5] described how genetic engineering tools can be used to solve instances of NP-complete combinatorial problems. Their work has led to hopes of a DNA computer that can outperform the fastest realizable super computers at such problems. Other suggested applications include massive associative memory [3] and problems in combinatorial chemistry, such as drug design [2].

Adleman solved a tiny instance of the Hamiltonian Path problem using a test tube-based methodology [1]. The chemical basis for solution of NP-complete problems is to represent the set of combinatorially possible solutions as a test tube of DNA strands (or oligonucleotides) and to selectively isolate the desired solutions from this set using hybridization, ligation, and other DNA manipulation processes. Problems exist with scale-up of this test tube-based approach for a number of reasons, including the poor efficiencies of the purification and separation steps.

An alternative methodology, in which the DNA strands are immobilized on a surface, offers many advantages for robust DNA computations. The solution set of oligos is initially attached to a surface (glass, silicon, gold, or beads, for example). They are then subjected to operations such as hybridization from solution or exonuclease degradation, in order to extract the desired solution. This method greatly reduces losses of DNA molecules during purification steps. We note that surface-based chemistries have become the standard for complex chemical syntheses such as solid-phase protein synthesis, solid-phase DNA synthesis, solid-phase protein sequence analysis, and many other chemistries [6]. The power of this method was also recognized by Adleman in his use of magnetic support particles in the Hamiltonian path experiment, although his approach did not fully exploit the utility of the support-based approach.

In this paper, we describe a "bare-bones" surface-based model for DNA computation. Our intent is to keep the model simple (incorporating only those operations that we can implement now), yet powerful enough to solve non-trivial exhaustive-search algorithms. We first describe (Section 2) an abstraction of the surface-based model. Briefly, the allowable operations are to selectively mark strands, to destroy either marked or unmarked strands, and to unmark all marked strands. We show that although this model may appear to be more restrictive than Lipton's model, it can be used to implement an exhaustive search algorithm for the Satisfiability problem. In Section 3, we describe in detail how the abstract model can be realized with standard surface-based chemistry. Another feature of our approach is the use of single-base mismatch discrimination in hybridization as a basis for selectively marking DNA strands immobilized on a surface. This method allows us to obtain a high density of information per base.

Partial experimental progress towards our initial goal of solving a 5-variable SAT instance using this model is described in Section 4. One experiment tests our method for selectively marking strands using hybridization. At this stage, we have not achieved a satisfactory level of success with this operation, and we discuss how we plan to improve on this. The second demonstrates successful use of the enzyme Exonuclease I to destroy single-stranded DNA. The third experiment involves both hybridization followed by enzymatic destruction. The results indicate that the hybridized strands are not destroyed, confirming that marking followed by destruction of unmarked strands is feasible.

It is not clear to what degree one can scale up the operations used to realize our surface-based model. In Section 5, we enumerate some of the limitations of our current approach, and suggest possible extensions that we would like to explore in the future.

2 Abstract Model

For simplicity, we let the initial solution space be the set S of binary strings of length n, that is, the set $\{0,1\}^n$. The following operations may be performed on S.

1. **mark**(i,b): this marks all strings of S in which the ith bit has value b.
2. **mark**$((i_1,b_1),(i_2,b_2),\ldots,(i_k,b_k))$: this is an extension of **mark**(i,b) in which a string is marked based on the values of many bits.
3. **destroy-marked**: removes all marked strings from the set S
4. **destroy-unmarked**: removes all unmarked strings from the set S
5. **unmark**: this unmarks all marked strings in S
6. **test-if-empty**: this operation determines whether the set S is empty or not. We assume it is only executed at the end of a computation.

The mark operation here is similar to the select operation of Lipton's model. However, our model is more restrictive than Lipton's, since marked and unmarked strings cannot be physically separated and handled differently. The only operation in our model that allows us to take advantage of marked strings is the destroy operation. Nevertheless, we now show that the Satisfiability (SAT) problem can be solved with these operations.

2.1 Algorithms for SAT

The SAT problem is defined as follows. We are given a Boolean formula in conjunctive normal form, over the set of variables $\{x_1,\ldots,x_n\}$. That is, the formula is the conjunction (logical "and") of clauses, each of which is the disjunction (logical "or") of some variables or their negations. The problem is to determine if such a formula is satisfiable; that is, if there is an assignment of truth values to the variables that sets all clauses to "true".

We present two different exhaustive search algorithms for the SAT problem. The solution space S represents all possible truth assignments, with 1= "true" and 0 = "false".

for each clause C of the form $(x_{i_1} \vee \ldots \vee x_{i_k} \vee \bar{x}_{j_1} \vee \ldots \vee \bar{x}_{j_l})$ do
    mark$((i_1,0),\ldots,(i_k,0),(j_1,1),\ldots,(j_l,1))$
    comment: all remaining solutions that set C to "false" are marked
    destroy-marked
test-if-empty

```
for each clause C do
    for each unnegated variable x_i in C do
        mark(i, 1)
    for each negated variable x_i in C do
        mark(i, 0)
    comment: all remaining solutions that set C to "true" are marked
    destroy-unmarked
test-if-empty
```

3 A Surface-Based Realization of the Abstract Model

We next describe the chemistry that we are currently developing to implement the above abstract model. We stress that some specific implementation choices that we are making now are not likely to scale to a solution space of size 2^{70}. Issues that arise in scaling up our current approach are discussed in Section 5.

3.1 Solution Space Representation

First, consider the simple problem of representing all possible binary strings of length n as DNA strands. Widely available methods for the solid phase synthesis of DNA molecules can be readily adapted to the synthesis of very complex combinatorial mixtures. In standard solid phase DNA synthesis, a desired DNA molecule is built up nucleotide by nucleotide on a support particle in sequential coupling steps. For example, a support with the nucleotide "A" attached may have the "A" reacted with a "C" to form a dimer, washed and the "C" coupled with "G" to form a trimer (still attached to the surface) and so on. This same chemistry can produce combinatorial sets of molecules by using mixtures of nucleotides at each coupling step. For example, if two nucleotides are used together in five coupling steps, 32 different molecules are made and are present on the support.

For now, we propose to use one base to represent one bit of the binary strings. In addition, all oligos representing a binary string have markers at each end; these will be used as primers in PCR reactions to obtain readout. We can use one base per bit while keeping the GC content of the string constant (say at 50%), by using A or T in half of the positions to represent 0 and 1 respectively, and C or G in the remaining positions to represent 0 and 1 respectively. This is an important aspect of the design of the combinatorial set, as GC content has a very strong effect upon hybridization stability and hence upon hybridization conditions.

3.2 Attachment Chemistry

The attachment chemistry describes the molecules at the interface of the surface and the oligos to be attached to the surface. (Both the surface and one end of the oligos are specially prepared to enable this attachment.) A good attachment chemistry ensures that the properly prepared oligos are immobilized to the surface at a high density, and that other oligos exposed to the surface later (for example, during hybridization) do not bind non-specifically to the surface.

We use a glass surface and an attachment chemistry developed in the Smith laboratory [4]. The glass surface is modified with amino-reactive isothiocyanate functionalities in a multi-step process.

3.3 Implementation of Operations

We now describe how each of the operations of our abstract model can be realized on surfaces.

mark(i, b): Let $S(i = b)$ be the set of binary strings of length n in which the ith bit is b. (There are 2^{n-1} strings in $S(i, b)$.) First, the set of DNA oligos that are complementary to the strands in $S(i = b)$ are synthesized. (This is done in the same manner that the initial solution space is synthesized, except that at the ith coupling only one nucleotide is introduced.) Then, each of these hybridizes to its complement on the surface (if present). Thus, the oligos to be marked are now double-stranded whereas those unmarked are single-stranded.

Note that this method assumes that a single-base mismatch is sufficient to discriminate between marked and unmarked oligos. In previous studies [4] it was determined that excellent specificity and discrimination of single-base mismatches is obtained using 15mer sequences. In Section 5, we address the fact that this becomes more difficult for longer oligos.

destroy-marked, destroy-unmarked: Either double-stranded or single-stranded DNA molecules may be selectively destroyed using enzymes known as exonucleases, which chew up DNA molecules from the end in, and exist with specificity for either the single-stranded or double-stranded form. We use the enzyme Exonuclease I to destroy single-stranded oligos.

unmark: This is done simply by washing the surface in distilled water. In the absence of salt, which stabilizes the double-stranded pairs, the complementary strands of DNA denature from the oligos on the surface and are washed away, leaving only the original single-stranded DNA attached to the surface.

test-if-empty: One way to implement this is to cleave the attached DNA from the surface, use PCR to amplify this and detect if there is any product as a result. Alternatively, if the strands on the surface are marked at the end of the computation, the complementary strands can be removed as described in the unmark operation and PCR can be applied to those. (In the experiment that we describe in Section 4, the design ensures that there is only one solution to the SAT instance, and so cycle sequencing will be used to actually read out the solution.)

4 Progress To Date

We are working on an implementation of the operations for a solution space of size 32. Our initial computational goal is to solve a 5-variable instance of SAT.

4.1 Solution Space Generation

The solution space, that is, set of 32 distinct molecules, which we have designed and synthesized have the sequences shown in Figure 1 below. They were synthesized using an Applied

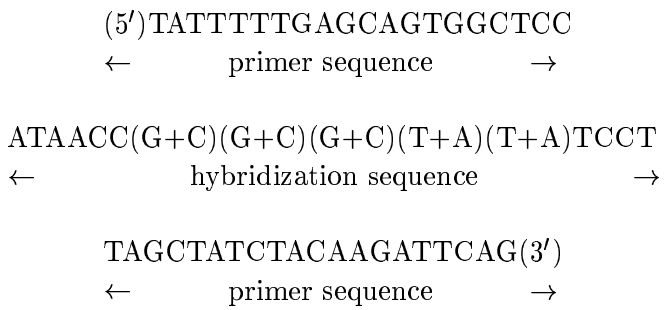

Figure 1: Representation of the Solution Space for a 5-Variable SAT Instance

Biosystems DNA Synthesizer.

These DNA molecules have an overall length of 55 nucleotides (nt), consisting of a unique 20 nt sequence at each end, and a 15 nt hybridization sequence in the middle. The 20 nt sequences permit both Polymerase Chain Reaction (PCR) amplification (by virtue of the two primer binding sites) of the combinatorial sets of DNA molecules and direct sequence analysis (by virtue of the primer extension site). The 15mer sequence in the middle was chosen to match the length of the oligonucleotides employed in the Smith laboratory for mutation detection [4]. The 5 central nucleotides in the 15 mer hybridization sequence were synthesized as a combinatorial set with two possibilities at each position: (G+C)(G+C)(G+C)(A+T)(A+T), yielding $2^5 = 32$ distinct products in the mixture. The adjacent 5 nucleotide sequences at the two ends of the hybridization sequence were synthesized as unique sequences here to limit the size of the combinatorial set to 32 molecules (later we will extend this work to a solution space of size 2^{15}). By keeping the number of positions in the 15mer encoded as A or T roughly equal to those encoded as G or C, the GC content is kept constant at about 50%.

4.2 Implementation of Operations

We next describe three studies that test the mark and destroy-unmarked operations defined above.

4.2.1 Marking by Hybridization

In this experiment, we followed previous work [4] which successfully achieved single-base mismatch discrimination in hybridization. We note that ultimately we will have do this experiment with the full solution space (32-mixture of molecules) on the surface, but for testing purposes we are currently using just two different molecules on the surface.

We attached two different single-stranded 5'-amino terminated DNA strands to a glass surface. The immobilized single-stranded DNA are both 30-mers with 15 T's as a spacer and the remaining 15 bases as hybridization sequence. The first strand (strand A) is complementary to the center 15-mer on one of the 55mers of our solution space. The other (strand B) is not complementary to any center 15-mer of the 32-mixture occurring in the solution space, and in fact there is a 1-base mismatch between the complement of strand B and some center strand

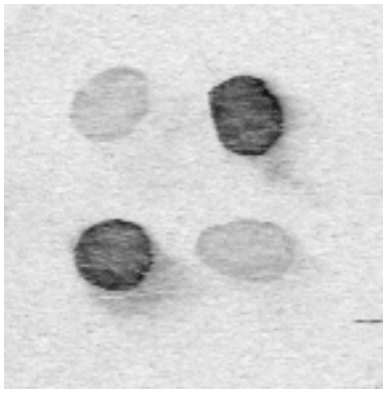

Figure 2: Marking by hybridization

in the solution space. The strands are attached in a group consisting of 4 spots. The two spots in one diagonal are both strand A and the two spots on the other diagonal are both strand B.

The solution space was fluorescently tagged and was hybridized with the A and B spots on the glass for 30 minutes at 37 degrees. Following this, the surface was washed in a buffer, in order to remove excess fluorescently tagged mixture and to improve flourescence detection of the hybridized solution space. Finally, the slide was scanned using a Molecular Dynamics FluorImager 575 for the detection of fluorescence. Hybridization patterns detected in this way show that hybridization occurred to strand A (two darker spots in diagonal) but only to a lesser extent to strand B (two pale spots in the other diagonal). The result is shown in Figure 2.

As can be seen from this figure, some unwanted fluorescence is detected in the mismatched spots. This could be due to (i) non-specific binding of strands on the surface, or (ii) bad discrimination, causing hybridization to occur even in the presence of a single-base mismatch. We have some evidence that (i) is a problem, since fluorescence is still detected after washing with water (which should break the hydrogen bonds of double-stranded DNA). In [4], bad discrimination was not a problem, but in their experiment, only a single fluorescently tagged DNA strand, rather than a mixture, was introduced to hybridize to the DNA strand immobilized on the surface. It is possible that the mixture of 32 distinct strands used in this experiment caused the increase in background. We are currently working on alternative surface chemistries to eliminate problem (i) and to increase the effectiveness of discrimination.

4.2.2 Destroy-unmarked: Destruction of Single Strands via Exonuclease I

We chose to selectively destroy single-stranded DNA molecules by using Exonuclease. Initially, Exonuclease I and VII were chosen to study the destruction of single strands immobilized on the surface. In preliminary studies, it was observed that Exonuclease VII not only had single-strand-directed $5' \rightarrow 3'$ and $3' \rightarrow 5'$ exonuclease activities, but also had some endonuclease activities (data not shown). Because of this, we chose not to use Exo VII as our enzyme, since endonuclease activities might accidentally destroy a valid solution. On the other hand, Exonuclease I acts specifically on single-stranded DNA, degrading it processively in the $3' \rightarrow 5'$ direction, thus producing $5'$ mononucleotides and leaving the terminal $5'$ dinucleotide intact [7].

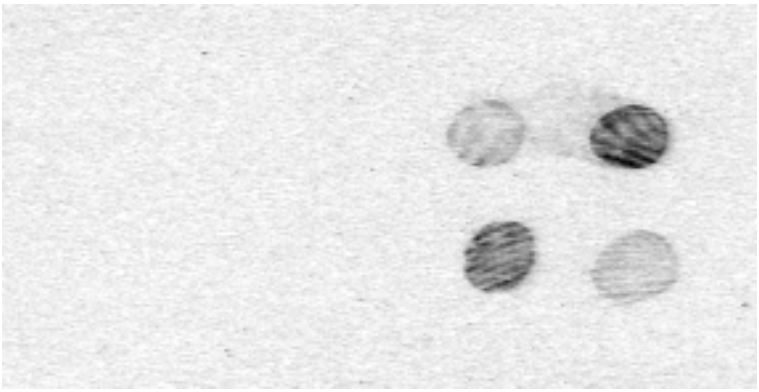

Figure 3: Destruction of single strands using Exonuclease I

In our experiments, we found that Exonuclease I could easily destroy single strands in solution without any added proteins or surfactants. However, when used to attack single strands attached to a surface, we found that BSA (Bovine Serum Albumin) or non-ionic surfactant Triton X-100 was needed to get the desired nuclease activity.

The experiment shown in Figure 3 demonstrates that Exonuclease I can destroy single-stranded DNA attached to the surface. Strands A and B were immobilized on the glass surface. 8 total spots were applied to the surface in two groups of 4 spots each. In each group, strands were spotted as described in the previous experiment. One group of 4 spots was the experimental group treated with Exonuclease I, while the other group of 4 spots was used as a control (no Exo I). After exonuclease digestion of the strands in the experimental group, the entire surface was hybridized with fluorescent complementary DNA strands. In the control group (right side of Figure 3), we should see the fluorescent signal, since immobilized single-stranded DNA was not destroyed and should hybridize to fluorescently tagged complement. In contrast, the group which was exposed to Exonuclease I (left side of Figure 3) should be destroyed and no fluorescent signal should be detected.

We conclude that Exonuclease I can destroy single-stranded oligonucleotides immobilized on the glass surface with the help of BSA or Triton X-100.

4.2.3 Destruction of Unmarked Strands Versus Marked Strands

In this experiment, after attaching strand A in two groups of 8 spots, 4 spots on the left were marked (hybridized) with their complementary strand in the flourescent 32-mixture, and 4 spots of A on the right were left unmarked (still single-stranded). Exonuclease I solution was added to both groups and incubated for 3 hours at 37 degrees. Then, the surface was washed to remove the enzyme and anything it might have digested. Next, the entire surface was hybridized to another portion of the fluorescent complement for 30 minutes at 37 degrees to reveal whether or not any strand A remained on the surface. After washing with washing buffer, the slide was scanned with the FluorImager. As shown in Figure 4, marked strands (on the left) appear to have been protected from destruction by Exonuclease, while unmarked strands (on the right) were destroyed by the enzyme.

Figure 4: Destruction of unmarked versus marked strands

5 Scaling up Surface-Based DNA Computation

In our approach, by immobilizing strands to a surface, loss of strands is greatly reduced. Also, our use of single-base mismatches to discriminate between strands to be marked and those to remain unmarked has important implications. First, a high information density - one base per bit or better - can be achieved. Second, this allows synthesis of solution spaces (such as $\{0,1\}^n$) quickly and cheaply using standard technology.

However, there are also limitations of the surface-based approach as described thus far. We enumerate these now, and suggest possible ways of overcoming these limitations.

The scale of computation is severely restricted by the 2-dimensional nature of surface-based computation. With the surfaces currently used, if each DNA molecule occupies an area of 20 angstroms by 20 angstroms, or 400 square angstroms (this is the density found to be optimum in our previous work, [4]), then a 1 cm area can accommodate only 2×10^{13} molecules. To go higher one must either a) increase the surface density, b) increase the surface area, or c) build linkage chemistry extending out into solution from which the oligonucleotides can be attached to make a local three-dimensional network on the surface.

Another limitation is that discrimination of single base mismatches in hybridization reactions becomes more and more difficult as the length of the oligos increases from the 15mers that we are now working with. One approach is to design the sequences such that they can be subdivided into independent "words" of, for example, 15 or 20 nt each. For example, a combinatorial set of molecules 60 nt in length would consist of four sequential 15 nt words. Targeting of each word would be accomplished with a complementary 15mer specific for the word being targeted. This places additional constraints upon the sequences, however, and further work is required to design the proper structure for the word elements.

Although our initial studies have employed glass surfaces, we do not anticipate that such glass microscope slides will prove to be the optimal surface for this application. We already noted that one problem with glass is non-specific binding of DNA molecules to the surface (that is, rather than a DNA molecule in solution hybridizing to its complement on the surface, it binds in some other manner to the surface itself). Thermally grown oxides on Silicon wafers or alkanethiol self-assembled monolayers on gold surfaces may be better alternatives.

Since hybridization efficiency is not 100%, some strands which should be marked may not actually be marked, and thus may be destroyed unintentionally. One solution to this problem is to add redundancy to the solution space. Another is to exclusively mark strands that should be destroyed, and use the destroy-marked operation. Now, the opposite problem may occur, in which strands that should be destroyed are not actually destroyed. However, this can be more easily corrected by repeating the mark and destroy-marked operation.

Finally, other operations should be added to the model, to increase its utility. Different methods for labeling desired subsets of the DNA molecules are needed that modify the desired DNA molecules based upon properties of their sequences. This modification must then allow selective manipulation of the labeled subpopulation in successive queries, so that the desired solutions may be found to the mathematical question posed. Many different labeling reactions are possible, based on a) exonuclease degradation of either single-stranded or double-stranded DNA; b) ligation of a marker oligonucleotide onto the ends of selected strands; c) polymerase extension from the ends of selected strands; and d) endonuclease cleavage of selected strands. With the ability to do b) for example, it would be possible to solve the Circuit Sat problem with the surface-based model.

6 Conclusions

We propose a new methodology, namely surface-based DNA computation, that offers many advantages over Adleman's approach. We focus on a very simple set of operations that are standard on surfaces, and show how the Satisfiability problem can be solved with these operations. We describe partial progress on experimental testing of these operations.

In its current form, our approach probably will not scale to manipulation of extremely large solution sets. In spite of this, we believe that further study of surface-based DNA computation will play an important role in the area of DNA computation. First, we predict that in the short term, computations involving several successive operations can be done at a greater scale on surfaces than in test tubes, and these can be used to learn much about the chemical processes underlying any DNA-based computational paradigm. Second, the techniques developed for our current model and the resulting improved understanding of the chemical processes used may well be useful in combinatorial chemistry applications, where the scale of experiments is much less than 2^{70}.

References

[1] Adleman, L. M. (1994). Molecular Computation of Solutions to Combinatorial Problems. Science, 266, 1021-1024.

[2] Adleman, L. M. (1995). On Constructing a Molecular Computer, Manuscript, Computer Science Department, University of Southern California.

[3] Eric Baum (1995), How to build an associative memory vastly larger than the brain, Science, April 28, 1995.

[4] Guo, Z., Guilfoyle, R. A., Thiel, A. J., Wang, R. and Smith, L. M. (1994). Direct Fluorescence Analysis of Genetic Polymorphisms by Hybridization with Oligonucleotide Arrays on Glass Supports. Nucl. Acids Res., 22, 5456-5465.

[5] Lipton, R. J. (1995). DNA Solution of Hard Computational Problems. Science, 268, 542-545.

[6] Smith, L. M. 1988. Automated synthesis and sequence analysis of biological macromolecules, Analytical Chemistry 60, 381A-390A.

[7] Lehman, I.R. and Nussbaum, A.L.(1964) J. Biol. Chem. 239, 2628-2636.