(45)

Design and Implementation of Computational Systems Based On Programmed Mutagenesis

by

Julia Khodor

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Science in Computer Science and Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 1998

Author .....
Department of Electrical Engineering and Computer Science
May 18, 1998

Certified by .....
David K. Gifford
Professor
Thesis Supervisor

Accepted by .....
Arthur C. Smith
Chairman, Department Committee on Graduate Students

Design and Implementation of Computational Systems Based On Programmed Mutagenesis

by

Julia Khodor

Submitted to the Department of Electrical Engineering and Computer Science
on May 18, 1998, in partial fulfillment of the
requirements for the degree of
Master of Science in Computer Science and Engineering

Abstract

We introduce *programmed mutagenesis*, a technique for rewriting DNA strands according to programmed rules. We present the Unary Counter– a simple computational system based on programmed mutagenesis.

We present the experimental results of the first two cycles of the unary counter which show that string rewriting by programmed mutagenesis is possible. We discuss the two enzyme system which allows to implement programmed mutagenesis by thermocycling a single reaction in a single test tube.

We discuss the sources of rewrite rule specificity within the framework of programmed mutagenesis and present experimental results which indicate that it is possible to guarantee the specificity of the rewrite rules, and thus the correctness of the computation.

We demonstrate that two oligonucleotides annealing to the template next to each other can ligate and extend to form the correct product. We argue that the ability to have oligonucleotides in close proximity function properly, together with the MIMD nature of the programmed mutagenesis systems provides evidence that parallel and nondeterministic computations are possible with programmed mutagenesis.

Finally, we discuss the significance of our results and outline directions for future work.

Thesis Supervisor: David K. Gifford
Title: Professor

Acknowledgments

The author would like to thank Professor David K. Gifford for his support and guidance through the duration of this research project and for his help in editing this document.

The author would like to thank Alexander J. Hartemink for utilizing SCAN [9] to search for the oligonucleotide primer sequences.

In addition, the author gratefully acknowledges the support of the Genomics Training Grant.

Contents

List of Figures

List of Tables

Chapter 1

Intoduction

Biological Computing has attracted interest since Adleman's original paper [2] because of its potential for high performance parallel computation. Lipton [12], Adleman [3], and others have proposed that the intrinsic power of processing large numbers (10^8) molecules in parallel may permit DNA computers to solve previously intractable problems. There are two primary approaches to DNA computing presently being pursued. The first set of approaches are based on generate-and-test for combinatorial search, and the second set of approaches seek to directly use DNA molecules to construct a universal computer.

1.1 DNA-based Methods for Combinatorial Search

Adleman introduced the combinatorial search approach [2], which relies on generating a set of witness DNA molecules and then searching among them for those that satisfy a given set of constraints. Each witness molecule encodes a single solution to the problem being solved.

Combinatorial search relies on the idea that the immense storage density of DNA allows the use of brute-force search methods to solve NP-complete problems. For example, Lipton [12] suggests using combinatorial search approach to break the DES

encryption code.

At present the manipulations required for DNA-based combinatorial search are complex and cumbersome. These techniques require human intervention between every step, which slows the computation down and introduces the potential for error. Despite optimism about future techniques, current techniques available for DNA-based combinatorial search are too slow and/or inaccurate to solve a problem of even moderate size.[10]

1.2 DNA-based Universal Computers

Several systems for constructing Turing machines have been suggested [4], [17], [21], [14]. The first two, by Beaver [4] and by Smith and Schweitzer [17], suffer from the same problems as the combinatorial search systems above. Namely, the biological techniques proposed for the implementation of these systems necessitate separating parts of the mixture and changing biochemical environments of the reactions during computation. This, again, both slows the computation down and introduces the potential for error. Winfree [21] and Reif [14] propose to implement universal computation via self-assembly of DNA. These techniques are based on the work by Seeman [16] exploring potential for constructing two- and three-dimensional structures out of a number of partially complimentary DNA molecules. Very few experiments studying the plausibility of this approach have actually been done to date. Preliminary results indicate that assembling even the simplest structures requires prolonged periods of time (16 hours or more for a 2-component system), with less then 50% efficiency of component incorporation [20]. There also exist so far unresolved concerns regarding the possibility that locally assembled substructures may interfere with the formation of global structures [14], [20].

Programmed mutagenesis is based on a string rewrite model of computation. DNA

naturally lends itself to a string rewrite model because the sequence of bases in DNA can directly be used to encode a string. DNA strand replication provides the ability to copy as well as the opportunity to introduce sequence specific changes into a newly synthesized molecule.

Since there is no known way to reliably mutate an existing DNA sequence, all DNA-based string-rewrite systems must include rewrite rules to be incorporated into the newly synthesized DNA strand. Thus, the main challenge to implementing any DNA-based string-rewrite system is guaranteeing the specificity of rewrite rules. This specificity is dependent on a particular implementation, as well as on the general restrictions posed by thermodynamics of the DNA-DNA and DNA-enzyme interactions.

There are several equivalent definitions of string rewrite systems [13]. Perhaps the most well-known string rewrite systems are Turing machines. A Turing machine is a tuple of the form (Q, Σ, δ, s), where

- Q is a set of states, including the start state s and the halt state h;
- Σ is a finite alphabet; and
- $\delta : Q \times \Sigma \rightarrow Q \times \Sigma \times \{L, R\}$ is a transition function.

A Universal Turing machine is a single Turing machine U, with the property that for each Turing machine T which computes a Turing-computable function f, there is a string of symbols d_T such that if the output of T on input x is $f(x)$, then the output of U on input $x\ d_T$ is also $f(x)$ [13].

A Turing machine is a string-rewrite system because operation of the machine can be described by a set of quintuplets of the form (*old state, symbol scanned, new state, symbol written, direction of motion*), i.e. as quintuplets in which the third, fourth, and fifth symbols are determined by the first and second. Thus, string rewrite systems are, in principle, universal [13].

Programmed mutagenesis is an in-vitro mutagenesis technique that uses DNA polymerase and DNA ligase to create copies of template molecules, where the copies have engineered mutations at sequence specific locations. Every time a programmed mutagenesis reaction is thermal cycled a rewriting event occurs. Because the technique relies on sequence specific rewriting, multiple rules can be present in a reaction at once, with only certain rules being active in a given rewriting cycle. Furthermore, the ability of the system to accommodate inactive rules allows the system to proceed without human intervention between cycles.

There are two main classes of possible designs for the DNA-based string-rewrite systems:

- Decoupled systems, where the sequences of the initial and final strings in each rewrite rule have no similarity or dependence relation between them; and
- Coupled systems, where the sequence of the final string of each successfully executed rewrite rule is strictly dependent on the sequence of the initial string being rewritten. Programmed Mutagenesis is an example of a coupled system.

Coupled and decoupled systems have different sources of specificity. Sources of specificity for both systems include the thermodynamics of DNA hybridization, and secondary structure considerations. In the coupled systems additional specificity comes from the relation between the sequences of strings being rewritten, and, equally important, from the relation between the sequences of the strings which should not interact. In particular, in programmed mutagenesis systems rewrite rule specificity is in part determined by the number and geometry of mismatches. The geometry of mismatches strongly influences the specificity of the rewrite rules. Other factors involved include temperature of the reaction, concentrations of templates and primers, and enzyme specificity.

All these parameters contribute to the dimensions of Hamming space around a given sequence. A *Hamming space* of a given oligonucleotide sequence s_1 is the section of

the sequence space around s_1 such that any sequence s_i within the Hamming space of s_1 can bind to and interact with s_1, and no sequence s_j outside of the Hamming space of s_1 can bind to and interact with s_1. Thus, programmed mutagenesis systems are Hamming rewrite systems. Figure 1-1A illustrates the definition of the Hamming space.

Accordingly, as illustrated in Figure 1-1B and C, allowed transitions are those that rewrite a string s_i to a string s_j iff the Hamming spaces of s_i and s_j intersect.

1.3 The Unary Counter

We have conducted a series of experiments to test the first cycles of a unary counter to explore the feasibility of the primitive operations required for programmed mutagenesis. In the context of our experimental work we have found that it is possible to create full-length product DNA molecules that have embedded rewriting, make later sequence changes to depend on earlier sequence changes, and have multiple oligonucleotides be active in close proximity on a template sequence.

In addition, our preliminary calculations show that programmed mutagenesis systems can operate at speeds of $\approx 4 \times 10^{13}$ polymerization operations per second, where an "operation" is polymerization of an entire DNA strand. These systems are also expected to exhibit storage density of $\approx 5 \times 10^{-6} um^3/bit$, where a bit is a computational symbol, not a base of a DNA duplex; and reduced power requirements of $\approx 10^{16}$ strand rewrite operations/J [7].

The structure of the unary counter system we are exploring is shown in Figure 1-2. As shown in the figure, oligonucleotide M-1 is responsible for creating a mutation in the first cycle product. This mutation permits oligonucleotide M-2 to bind in the second cycle, which in turn permits oligonucleotide M-1 to bind in a new location in the third

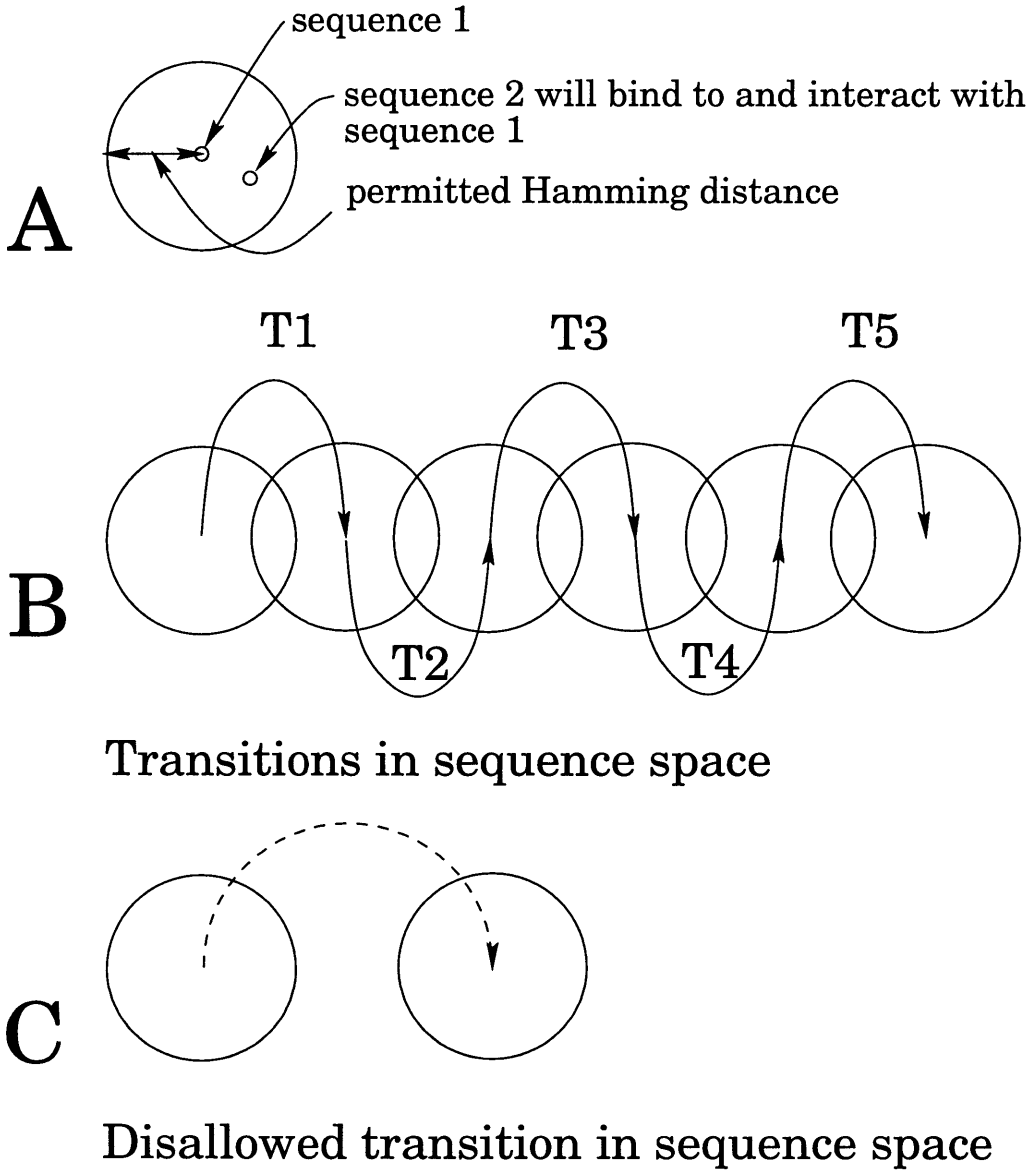

Figure 1-1: Schematic representation of Hamming rewrite systems.
Part A shows Hamming space of the sequence 1, including sequence 2 in that Hamming space, which will bind to and interact with sequence 1.
Part B illustrates allowed transitions in sequence space. T1 through T5 are allowed because they accomplish transitions between sequences whose Hamming spaces intersect. Part C illustrates a disallowed transition in sequence space. The Hamming spaces of the sequences in the figure do not intersect.

cycle. An "outer" oligonucleotide primer is used to create full-length products and the strand that polymerizes from the outer primer is ligated to a mutagenic primer by ligase. As described below, all of the enzymes used in the system are thermostable which allows the system to be thermal cycled.

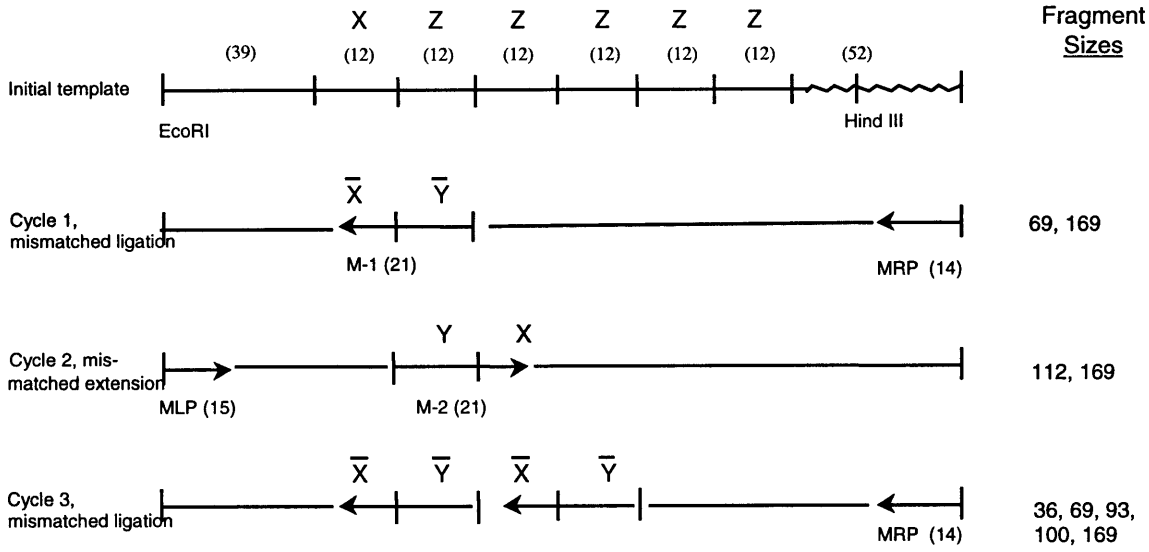

Figure 1-2: Schematic representation of the unary counter. M-1 and M-2 are mutagenic rule oligonucleotides; MRP and MLP are perfectly matched outside oligonucleotides. Note that a rule incorporated on a previous cycle becomes part of the template for the following cycle.

The DNA sequences for the system we are testing have been selected using the SCAN program [9] to search a large sequence space constrained by chosen mismatch geometry. SCAN chooses sequences that have optimum annealing properties, do not have harmful secondary structure, and do not form primer dimers.

In the remainder of this thesis we discuss data which demonstrates that string rewriting by programmed mutagenesis works (Chapter 2), demonstrate how to guarantee the correctness of the computation (Chapter 3), show that parallel and nondeterministic computations are possible with programmed mutagenesis (Chapter 4), and conclude with observations about the practicality of programmed mutagenesis for computation and directions for future work (Chapter 5).

Chapter 2

String Rewriting by Programmed Mutagenesis is Possible

The basic step of a string-rewrite computational system is the creation of a new string based on a template string. In programmed mutagenesis systems this step is accomplished by incorporating a mismatched oligonucleotide into the newly-synthesized DNA strand. Creating a new strand in a programmed mutagenesis reaction requires two enzymes. DNA polymerase is used to extend the strands and DNA ligase is used to join together parts of a new strand.

In this chapter we discuss how to create a two enzyme system which permits programmed mutagenesis to proceed in a single reaction. We then present data which indicates that the system works new strings can be created with embedded rewritings without undesirable artifacts.

2.1 Enzyme and Buffer Choice

In its native configuration DNA is a double helix, where the two strands are joined together by the hydrogen bonds between the bases (A, C, G, and T). Bases form stable pairs A-T and G-C. All other base pairings are considered mismatches. Mismatches

differ in stability, but all are less stable then the perfect match. DNA strands have polarity. The phosphate-sugar backbone to which the bases are attached is directed from the $5'$ to the $3'$ position of the sugar ring. New bases are added to the $3'$ end of the strand and two strands in the double helix are antiparallel.

The strict A-T, G-C pairings provide the basis for duplicating a DNA strand, since by looking at one strand we can reproduce the sequence of its compliment. DNA replication requires a starting point called a "primer" that is a short DNA molecule which binds to the old strand and creates a stable $3'$ end from which the synthesis of the new strand can proceed.

While living cells employ complex enzymatic machinery to "open" DNA helix in order to begin replication, we use thermal denaturation to achieve the same goal. Unfortunately, most enzymes are thermally deactivated at the temperature that is required for DNA denaturation. Therefore, in order to avoid having to add enzymes during each step of the computation we employ thermostable enzymes.

We have developed an enzyme system that permits a mutagenic oligonucleotide to be embedded in newly synthesized DNA strand. As shown in the first cycle of Figure 1-2, in this system a mutagenic oligonucleotide serves as a primer for DNA polymerase on its $3'$ end, and accepts a DNA ligation event on its $5'$ end. In our system these events occur in the same reaction at the same temperature.

The two enzymes that we use in our system are Taq Ligase and Vent Polymerase. Taq Ligase [5] has the virtue of being the only commercially available thermostable ligase. Vent exo^+ Polymerase [5] does not have $5' \rightarrow 3'$ exonuclease activity, and does not unacceptably strand displace at 45^oC. In order to prevent $3' \rightarrow 5'$ proofreading of mutagenic oligonucleotides we manufacture these with sulfur instead of phosphorus linkages on the last four bases. These phosphothioate linkages render the oligonucleotide extendible, but not degradable. Vent exo^- Polymerase is similar to Vent,

but lacks the $3' \rightarrow 5'$ proofreading function and strand displaces more then Vent exo^+.

In order for a two-enzyme system to function, both enzymes must function efficiently in a single buffer. Taq DNA Ligase requires NAD as a cofactor. We constructed a custom buffer by adding 10 mM of NAD to 10X Thermopol Vent buffer (hence forth called Vent-NAD buffer). We then tested the efficiency of Vent exo+ and exo- DNA polymerases and Taq DNA Ligase in Thermopol buffer, Vent-NAD buffer and Taq Ligase buffer. Taq Ligase buffer and Vent-NAD buffer allow 100% ligation of the control Bst I cut lambda DNA, while Thermopol buffer alone allows only incomplete ligation. However, Taq Ligase buffer does not support efficient polymerization, and thus we chose 10X Vent-NAD buffer for all further experiments.

We hypothesized that reducing the rate of strand extension would help increase the probability of successful ligation events in our system. A molecule of Vent DNA Polymerase extends DNA in increments of 6-7 bases at a time [11]. It then "falls off" the template, and searches for another open $3'$ end to extend. Molecules of Taq DNA Ligase are also constantly searching for a suitable target. We theoretized that once polymerase encounters a properly aligned downstream oligonucleotide, there would be competition between the two enzymes for the ligation site. We thought that it is possible that even though the polymerase cannot extend the $3'$ end any further, it still recognizes the $3'$ end as a possible site of action. Thus having an unligated $3'$end next to another strand may enhance the ability of Vent polymerases to displace the other strand.

We experimentally found that excessive amounts of polymerase reduced the amount of ligation product in our system. We tested four different concentrations of the polymerase (1U, 0.25U, 0.125U, and 0.05U per reaction) and concluded that the highest efficiency is achieved with 0.25U of polymerase and 40U of ligase per 10μl reaction (data not shown).

2.2 Complete molecules with intended rewriting events are produced

To demonstrate that DNA molecules can be created with internal rewriting events we designed an experiment based on the first cycle reaction of the unary counter (Figure 1-2). This system consists of a perfectly matched upstream oligonucleotide MRP and a mismatched downstream oligonucleotide M-1. The result of running such a system at four different temperatures (42.5, 45, 47.5, and 50°C) is displayed in Figure 2-1 (lanes 1, 2, 7, 8, 13, 14, 19, 20). In all reactions the M-1 oligonucleotide has been end-labeled with 32P, and the creation of 169 bp product demonstrates that both polymerization and ligation has occurred. The 69 bp template is extension product that did not participate in a ligation reaction.

The $10\mu l$ reactions above included 2 x 10^{-13} moles of the initial template molecules, 2 x 10^{-12} moles of the outside primer MRP, and 2 x 10^{-13} moles of the M-1 oligonucleotide. Reactions were first denatured at 94C for 10 minutes, and then brought down to the appropriate reaction temperature (42.5, 45, 47.5, or 50°C) for 30 minutes. The reactions were stopped by adding $10\mu l$ of Stop/Loading Dye, which stops all DNA interactions from proceeding further.

In addition to the appearance of the expected products, this experiment also confirmed that products related to non-specific primer binding did not occur.In this experiment the ZZ region of the template offers binding spots that contain four mismatches instead of the two mismatches found at the correct binding site. Although there is a 20^oC difference in the predicted melting temperature between the desired binding location ($T_m = 45^oC$) and incorrect locations ($T_m = 25^oC$) we still wanted to ensure that no inappropriate binding occurs. As shown in Figure 2-1, no products

are produced at the four mismatch binding sites.

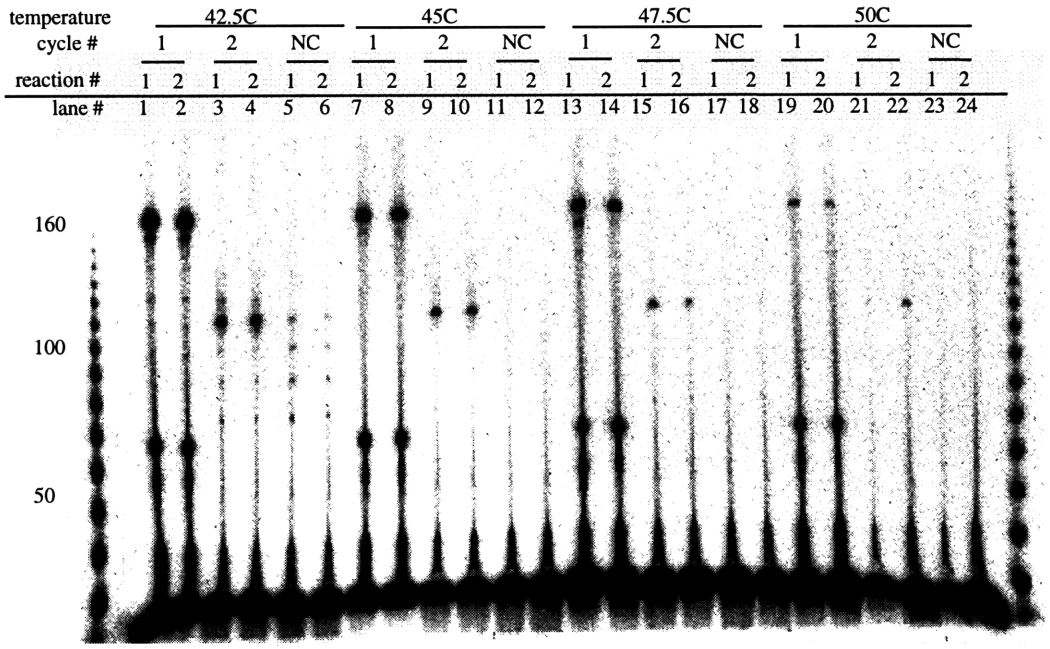

Figure 2-1: Temperature dependence of first and second cycle reactions from Figure 1-2. In each temperature range the first two lanes represent cycle1, the next two cycle two, and the last two negative controls. Note that MLP is absent from all reactions. Thus we do not expect bands larger then 112bp in the second cycle lanes. In the first cycle reactions M-1 is end-labeled by ^{32}P. In the second cycle and negative control reactions M-2 is end-labeled by ^{32}P.

In our experimental system Taq DNA Ligase requires 30 minutes to achieve optimal ligation efficiency. As a point of reference, Vent DNA Polymerase requires only about a minute to complete polymerization of the full-length product. As the proper binding sites for the rule oligonucleotides are depleted, effective conditions in the system change. The concentrations of free rule oligonucleotides are still high, while the effective concentration of proper binding sites decreases. As the number of proper binding sites diminishes, improper sites become more favorable (as compared to the option of staying in the single-stranded state). As a result, the system's drive toward equilibrium forces even unfavorable bindings to take place. Thus, when designing bio-

logical computing systems, we need to carefully consider the influence of such factors as primer and template concentrations, reaction temperature, reaction volume and reaction time on the T_m of oligonucleotide rules with respect to proper and improper binding sites.

Another family of artifacts we need to be concerned about are primer-dimers. A primer-dimer is a complex of oligonucleotides partially joined by base complimentarity. If the complimentarity is over the $5'$ region of the oligonucleotides, primer-dimers may bind free oligonucleotides into partially double stranded structures. If the complimentarity is over the $3'$ region of the oligonucleotides, the partially double stranded structures may be extended to copy the noncomplimentary region of the one primer onto another.

Oligonucleotide rules in the unary counter are complimentary to each other on their $3'$ ends. Under the favorable thermodynamic conditions they can bind to each other and extend, forming 36mer oligonucleotides. These, in turn can bind to each other and extend, producing 60mer polynucleotides. These dimer products may interfere with the reaction by producing full length products by "skipping steps," i.e. by a singular rewriting event rather then a series of sequence specific rewriting events. Furthermore, fidelity of computation is compromised since the longer polynucleotides have different thermodynamic characteristics with respect to the template, and may, therefore, potentially be able to rewrite the sequence such that the result is not a valid result of the intended computation.

Figure 2-1 shows that at temperatures in excess of 42.5°C, binding of our primers to sites other then the intended site of mutagenesis (non-specific binding) does not occur. There is also no observable primer dimer formation at any temperature.

We have demonstrated that primitive operations of programmed mutagenesis, extension and ligation of a mismatched oligonucleotide primer, can occur in a single buffer

at a single temperature. In Chapter 3 we discuss how to guarantee the correctness of a programmed mutagenesis computation.

Chapter 3

Guaranteeing the Correctness of Computation

3.1 Importance of programmed sequential incorporation

The fidelity of a Programmed Mutagenesis computation depends on the proper incorporation of mutagenic oligonucleotides. Suppose in a given computation the question posed is "Does this computation result in s3?" Now suppose that the set of rewrite rules for this computation includes s2 $\rightarrow$ s3, but not s1 $\rightarrow$ s3, or any other rules which allow s1 to be rewritten. Suppose further that the correct computation terminates with the incorporation of an oligonucleotide rule that rewrites the string into s1. Thus, the answer to the question posed for this computation should be "no." However, if s1 presents a suitable biological template for the execution of s2 $\rightarrow$ s3 rule, this rewrite rule will be incorporated, and it would appear that the result of the computation is "yes." In other words, if the rules are not incorporated strictly sequentially, no claims about correctness of computation can be made.

In case of the unary counter, in order to establish programmed sequential incorporation, we need to show that the second cycle product appears iff the first cycle has

executed successfully. In the rest of this chapter we discuss the sources of specificity of the rewrite rules and present results which indicate that it is possible to guarantee the correctness of a programmed mutagenesis computation.

3.2 Sources of rewrite rule specificity

As described in Chapter 1, specificity of the rewrite rules is determined in part by the Hamming distance constraints in the system. In addition to mismatch distance constraints, other sources of specificity are mismatch geometry, processivity of the enzymes, and thermodynamic parameters such as relative amounts and concentrations of template and primers in the reaction, salt concentration, time reaction is allowed to proceed, and reaction temperature. When we began our research, there existed reliable information about the thermodynamics of a perfectly matched DNA duplexes [6], [15], but information about the thermodynamics of mismatched DNA duplexes was less reliable [1], [18]. Through our early experiments we have acquired empiric data and developed intuition regarding mismatched duplexes.

With the appearance of BIND [8], we were able to relate our empiric observations and experiment results to a theoretical prediction. For instance, BIND determined that in our early experiments there was a very narrow T_m difference between oligonucleotides binding in a correct spot and those binding inappropriately. This explained the inappropriate products we were observing in that system.

As described in Chapter 1, we used SCAN [9] to choose particular sequences for the Unary Counter. We chose fairly strict thermodynamic constraints in order to prevent inappropriate binding of the rule oligonucleotides, as well as any undesirable interaction between primers. Nevertheless, the search space remained too large, and needed to be further constrained.

The greatest constraint on the oligonucleotide rules search space is placed by the choice of the geometry of mismatches. Mismatches drastically change the thermodynamic characteristics of the oligonucleotides. The characteristics depend on:

1. The particular mismatch used. For example, a C-A mismatch greatly strains the helical structure around itself, while a G-T mismatch has almost no effect on neighboring DNA structure [22].

2. The base pairs surrounding the mismatch [19], [15].

3. The position of the mismatch within the oligonucleotide. Mismatch too close to the 5′ end could destabilize ligation, while one too close to the 3′ end may disturb extension

4. The positions of mismatches with respect to each other. It is reasonable to expect that two mismatches right next to each other would have less of an effect on the stability of the duplex then would those same two mismatches located some distance away from each other.

The third point deserves further explanation. Polymerase and ligase enzymes have their own requirements for how stable the site of action needs to be in order for the enzyme to catalyze the reaction there. In our particular system, Taq DNA Ligase is extremely discriminatory with respect to the required stability of its site of action. Mismatches within the first four bases of the 5′ end of the primer being ligated to are not tolerated. Vent DNA Polymerase allows mismatches that are as close to the 3′ end as three bases. However, these are only tolerated if there is no other mismatch in the immediate vicinity.

Having spent much time in the laboratory working with our early system, we acquired certain intuition regarding mismatch geometry. Based on that, we proposed a design. We also investigated three of its close relatives.

To test our intuition and to collect more empirical data we designed a number of oligonucleotides. These primers were homologous to the elements from our early system and had mismatch structure and several other characteristics of each of the proposed mismatch designs. We used these primers to test the putative extension and ligation efficiencies of the proposed designs. We then used SCAN [9] to search for optimal oligonucleotides with the given mismatch geometry and a common set of thermodynamic constraints. It is interesting to note that the first design we proposed is the one we implemented in the laboratory.

3.3 Experimental results

In order to demonstrate the specificity of the rewrite rules, we need to show that second cycle oligonucleotide M-2 binds in the appropriate location when M-1 is present in the reaction, and not at all when there is no M-1 in the reaction. Figure 2-1 shows that M-2 product is only generated in cycle 2 when M-1 is present in the cycle 1. In lanes 3, 4, 9, 10, 15, 16, 21, and 22, an M-2 band of the appropriate size (112bp) is observed, and the intensity diminishes with an increase in temperature as would be expected with M-2's predicted melting temperature of 45^{o}C. As expected, there are no bands in the negative control lanes (11, 12, 17, 18, 23, and 24) at higher temperatures. There is some non-specific binding at the lowest temperature (42.5^{o}C) in the negative control lanes. This non-specific binding is completely eliminated by raising the reaction temperature by 2.5^{o}C.

We analyzed the data in Figure 2-1 by counting the amount of ^{32}P label in every band of the gel. Calculated efficiencies of the first and second cycles of the unary

counter operations are presented in Table 3.1. The plot of the reaction temperature vs. the amount of full-length product of first cycle is shown in Figure 3-1 and the plot of the reaction temperature vs. the efficiency of second cycle is shown in Figure 3-2.

The linear reduction of full length product with an increase in temperature in Figure 3-1 is quite striking. In our future experiments we hope to elucidate whether this dependency is general for all programmed mutagenesis systems, or particular to this design of the unary counter. We hope that this line of experiments will prove fruitful in our quest for understanding of the underlying principles of mismatch biochemistry.

Figure 3-2 shows that the highest efficiency of the second cycle incorporation is achieved at T=45°C. This result is expected since the predicted T_m of the M-2 oligonucleotide rule in the correct alignment is 45°C. At first glance such discrepancy in efficiencies between the first and second cycles seems baffling. However, if we consider the difference between the kinds of templates available on these cycles, we begin to appreciate the cause of this discrepancy.

T°C	%incorporated on cycle 1	%full-lenght on cycle 1	%ligation efficiency	%incorporated on cycle 2	%2nd cycle efficiency
42.5	9.82	5.25	53.41	2.48	47.19
45	8.55	3.64	42.51	1.8	49.52
47.5	5.73	1.88	32.81	0.62	32.71
50	5.37	0.34	6.24	0.5	error

Table 3.1: Quantification of the results from Figure 2-1. Second cycle efficiency is defined as the % incorporated on cycle 2 over the % full-length on cycle 1.

The first cycle unary counter template is embedded in a 3,000 bases long double stranded DNA vector. Thus, M-1 is competing for its binding spot with a 3Kb perfectly complementary strand. In order to produce the full-length product on the first cycle, both MRP and M-1 must bind to the same template strand, extend, and ligate together. With this in mind, 3.6% does not seem to be unreasonable. The full-length product of the first cycle reaction is a single stranded polynucleotide 169 bp long.

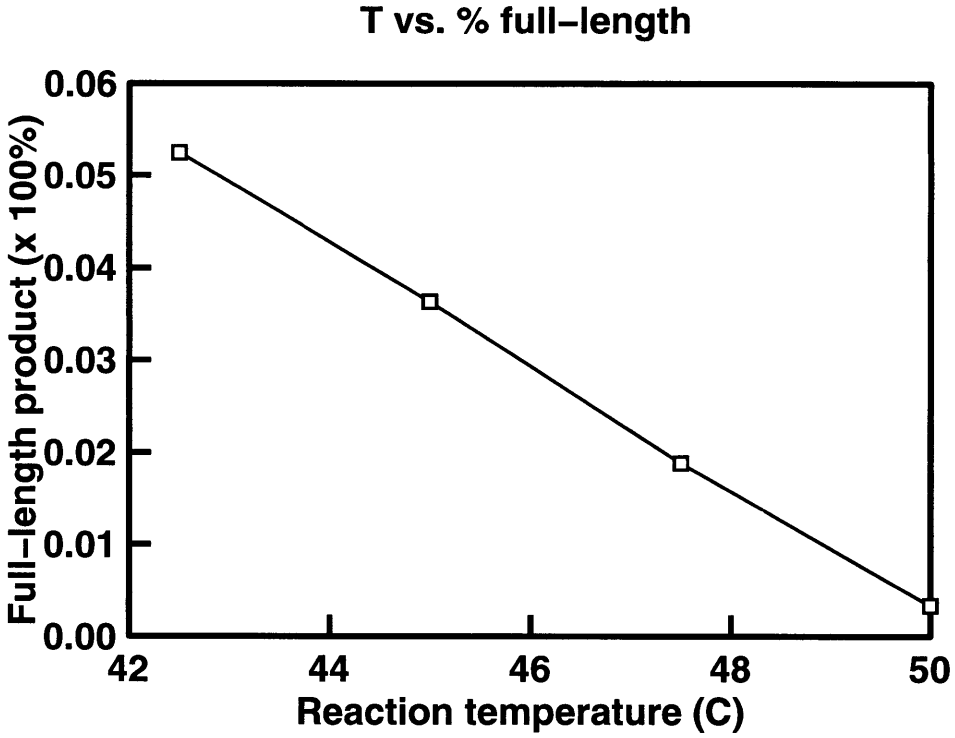

Figure 3-1: The plot of reaction temperature vs. % of full-length product produced in cycle 1 of the experiment in Figure 2-1.

Since the original template provides little competition to M-2, it is reasonable to expect an increase in the efficiency of the reaction. The efficiency of the second cycle reaction, 50%, represents a substantial increase over the efficiency of the first cycle reactions, and provides the reason to expect advanced cycles of the machine to run with a double-digit efficiency.

Altogether, Figure 2-1 demonstrates two essential functionalities of the Programmed Mutagenesis systems. It demonstrates both the specificity of each rewrite rule and the strict dependence of the activity of the next rule on the incorporation of the previous. In the next chapter we discuss the experiments which show that it is possible to incorporate more than one rewrite rule on a given cycle.

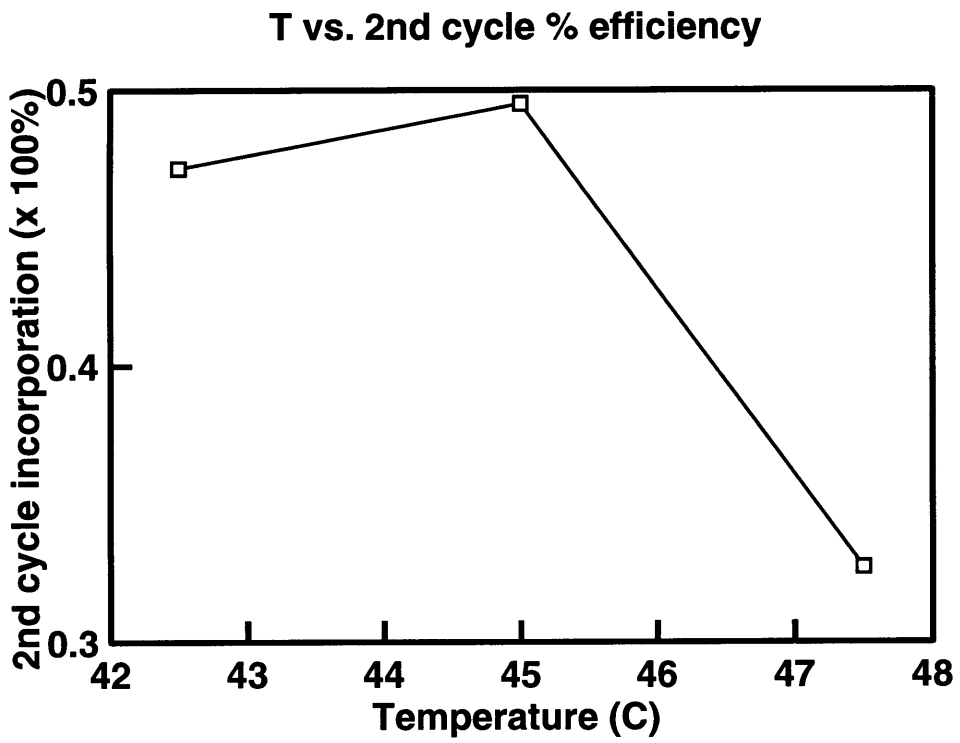

Figure 3-2: The plot of reaction temperature vs. % efficiency of the incorporation of the mutagenic oligonucleotide M-2 on the second cycle of the experiment in Figure 2-1. Second cycle efficiency is defined as the % of mutagenic oligonucleotide M-2 incorporated on cycle 2 over the % full-length product produced on cycle 1.

Chapter 4

Evidence That Parallel and Nondeterministic Computations are Possible With Programmed Mutagenesis

In order to demonstrate the flexibility of programmed mutagenesis we need to show that two oligonucleotide rules positioned on the template next to each other can ligate together and extend to the end of the template. This configuration of oligonucleotides is expected to occur when one oligonucleotide binds to the previously rewritten section of the template, while another binds to an adjacent mismatched sequence to be rewritten. It is reasonable to suspect that if the mismatched oligonucleotide is downstream of the perfectly matched one, the former may anneal earlier and extend over the binding site of the latter, thus preventing the latter from working.

As shown in Figure 4-1, our system consists of two 24 bases long oligonucleotides with immediately adjacent binding sites. The upstream oligonucleotide UP_CB is perfectly matched, while the downstream oligonucleotide CB has two mismatches with the template. We need to confirm that primer UP_CB can ligate to primer CB,

and that primer CB can be extended by a polymerase.

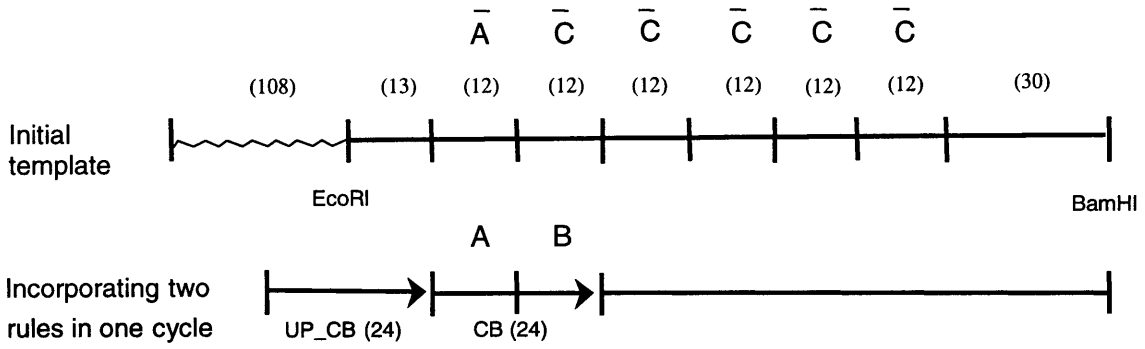

Figure 4-1: Schematic representation of UP_CB system. UP_CB is a perfectly matched 24b long oligonucleotide. CB is a 24b long mutagenic primer, which has two mismatches with the template when aligned as above.

We experimentally validated that the system in Figure 4-1 produced full-length products that included both of the oligonucleotides in the reaction (data not shown). Our design uses a 24 base long UP_CB oligonucleotide complementary to the region immediately upstream of the site of the 24 base long CB oligonucleotide. The test showed that UP_CB can ligate to the downstream mismatched oligonucleotide CB, albeit with a slightly lower efficiency than in the case of two oligonucleotides whose binding sites are separated by some distance. We also confirmed that adding UP_CB five minutes after the start of the reaction increased the ligation efficiency, and this effect was enhanced as the temperature of the reaction increased. We speculate that the polymerase becomes more active at higher temperatures, filling the bases over the binding site of CB before the latter could anneal.

The case we consider in Figure 4-1 is the worst case scenario of a multiple oligonucleotides being incorporated on the same cycle. We have already established that oligonucleotides some distance away from each other can extend and ligate together (Figure 1-2, MRP and M-1 oligonucleotides). Based on these results, it is reasonable to expect that two mismatched oligonucleotides some distance away from each other

can be incorporated into the new strand on a single cycle. Thus, we speculate that it is possible to incorporate a number of oligonucleotides, whether perfectly or imperfectly matched, on the same replication cycle.

Programmed mutagenesis is unique in its ability to coordinate the activities of independent computations in a single reaction because DNA is used for both program and data storage. This duality is a crucial aspect of programmed mutagenesis, and provides the basis for its Multiple-Instruction Multiple-Data (MIMD) architecture. With this in mind, it is easy to see that programmed mutagenesis can allow independent computational processes in a single reaction to communicate with each other by exchanging DNA. Each independent computational process in a reaction is represented by a unique template state, and templates are used to direct the manufacture of novel DNA sequences. Because one template can direct the manufacture of a DNA molecule that can interact with a second template, it is, in principle, possible to implement general communication between processes.

For example, once a counter increments to a preset value, it can produce a new DNA molecule with a "messenger" sequence that can be used to interact with a second template in subsequent computational cycles. The production of a messenger DNA molecule can be accomplished by choosing a primer for the first computation that will only be active once the counter has mutated its binding site. Once the counter counts to the preset value, and the primer binds, the polymerization of the messenger molecule can occur, and the messenger can be designed to directly interact with other templates in the reaction. It is also conceivable that new rule oligonucleotides can be generated as a programmed mutagenesis reaction proceeds, allowing program flow to be altered.

We have described the theoretical basis for the possibility of having a number of communicating computation proceed in a single programmed mutagenesis reaction. We have also completed experiments which demonstrate that it is possible to incorporate

more than one oligonucleotide into a single new molecule in a single cycle. Taken together, these two points suggest that parallel and nondeterministic computations may be possible with programmed mutagenesis.

Chapter 5

Programmed Mutagenesis is Viable

We have demonstrated that the number of key aspects of programmed mutagenesis approach function properly. In particular, we have demonstrated that string rewriting by programmed mutagenesis in a single buffer at a single temperature is possible; that incorporation of each oligonucleotide rule can proceed iff the previous rule was incorporated correctly; and that it is possible to incorporate more than one oligonucleotide rule into a single new DNA strand on a single cycle. We expect that we will be able to use programmed mutagenesis to implement a wide variety of computational primitives.

It is important to note that the applications of programmed mutagenesis technique do not end with biological computing. In fact, methods we developed would be useful in basic biological research for applications such as creating targeted mutations for basic genetics studies, drug design, gene therapy, in-vitro evolution, and a number of others.

More work needs to be done in order to simplify the design process for programmed mutagenesis systems. One of the most limiting factors in the present design of the unary counter is the choice of enzymes and the restrictions it poses on the number and geometry of mismatches allowed. While we were not able to find a better combination

of enzymes, it is possible that one can be found, as new and modified thermostable enzymes are introduced every year.

Another possible approach to easing enzyme-posed limitations on the system is to optimize the reaction buffer. While we have found a buffer which allows both enzymes to work efficiently as compared to their native conditions, it may be possible to further optimize the buffer so as to improve the ligation efficiency, and, thus, decrease the cycle time.

In order to perfect our understanding of the influence the mismatches have on the biochemical characteristics of the primers, better data on the mismatch biochemistry is needed. Data on the influence of the mismatches on the nearest-neighbor interactions in the DNA duplex is of particular importance. Better data on the mismatch biochemistry would allow us to refine the computer tools and, as a result, to simplify the design process.

We have empirically observed that the geometry of mismatches is the single most important element of the design of the programmed mutagenesis systems. However, at present our understanding of the mechanisms involved is mostly intuitive and anecdotal. We believe a large study of the influence of the geometry of the mismatches on the biochemical characteristics of the DNA duplex is in order. The study should endeavor to elucidate:

1. The precise relationship between the location of the mismatch relative to the site of action of an enzyme and enzyme's efficiency;

2. The comparative degree of instability introduced by the same mismatches depending on their context and distance from each other and the ends of the oligonucleotide; and

3. Whether chemically modifying the oligonucleotides, such as replacing phosphorus linkages by sulfur linkages, changes their biochemical characteristics in general, and in the particular case where some of the mismatches are located in the modified region of the oligonucleotide.

While the results of all the above studies would immensely simplify the design process for the programmed mutagenesis systems, we have demonstrated that it is possible to create the functional programmed mutagenesis systems.

Bibliography

[1] F. Aboul-ela, D. Koh, and I. Tinoco, Jr. Base-base mismatches. Thermodynamics of double helix formation for dCA3XA3G + dCT3YT3G (X, Y = A, C, G, T). *Nucleic Acids Research*, 13(13):4813–4823, 1985.

[2] L. M. Adleman. Molecular computation of solutions to combinatorial problems. *Science*, 266(5187):1021–1024, 1994.

[3] L.M. Adleman. On constructing a molecular computer. Technical report, UCLA, Department of Computer Science, UCLA, CA, January 1995.

[4] D. Beaver. Molecular computing. Technical Report CSE-95-001, Penn State University, Department of Computer Science and Engineering, Penn State University, PA, January 1995.

[5] New England Biolabs. *Biological research products catalog.* NEBiolabs, Inc., 1997.

[6] K.J. Breslauer, H. Blocker, Frank, and L.A. R., Marky. Predicting dna duplex stability from the base sequence. *PANS USA*, 83(11):3746–3750, 1986.

[7] D.K. Gifford, PI. Programmed mutagenesis and receptor-based sensor cells. *Proposal to DARPA ITO BAA#98-11*, 1997.

[8] A.J. Hartemink and D.K. Gifford. Thermodynamic simulation of deoxyoligonucleotide hybridization for dna computation. *Proceedings of the Third DIMACS Workshop on DNA Based Computers*, 1997.

[9] A.J. Hartemink, J. Khodor, and D.K. Gifford. Automated constraint-governed nucleotide sequence selection for dna computation. *Proceedings of the Fourth DIMACS Workshop on DNA Based Computers*, 1998.

[10] J. Khodor and D.K. Gifford. The efficiency of sequence-specific separation of dna mixtures for biological computing. *Proceedings of the Third DIMACS Workshop on DNA Based Computers*, 1997.

[11] H.M. Kong, R.B. Kucera, and W.E. Jack. Characterization of a dna-polymerase from the hyperthermophile archaea thermococcus-litoralis - vent dna-polymerasee, steady-state kinetics, thermal-stability, processivity, strand displacement, and exonuclease activities. *Journal of Biological Chemistry*, 268:1965–1975, 1993.

[12] R.J. Lipton. Dna solutions of hard computational problems. *Science*, 268:542–545, 1995.

[13] M.L. Minsky. *Computation: Finite and Infinite Machines.* Prentice-Hall, Inc., 1975.

[14] J. Reif. Local parallel biomolecular computation. *Proceedings of the Third DIMACS Workshop on DNA Based Computers*, 1997.

[15] J. SantaLucia, Jr., H.T. Allawi, and P.A. Seneviratne. Improved nearest-neighbor parameters for predicting dna duplex stability. *Biochemistry*, 35(11):3555–3562, 1996.

[16] N.C. Seeman. Denovo design of sequences for nucleic-acid structural engineering. *Journal of Biomolecular Structure & Dynamics*, 8(3):573–581, 1990.

[17] W.D. Smith and A. Schweitzer. Dna computers in vitro and vivo. Technical report, NEC Research Institute, NEC Research Institute, Princeton, NJ, March 1995.

[18] H. Werntges, H.J. Fritz, D. Reisner, and G. Steger. Mismatches in dna double strands– thermodynamic parameters and their correlation to repair efficiencies. *Nucleic Acids Research*, 14(9):3773–3790, 1986.

[19] J.G. Wetmur. Dna probes– applications of the principles of nucleic acid hybridization. *Critical Reviews in Biochemistry and Molecular Biology*, 26(3-4):227–259, 1991.

[20] E. Winfree. Algorithmic self-assembly of dna: Theory and experiment. Talk at MIT, April 1998.

[21] E. Winfree, X. Yang, and N. Seeman. Universal computation via self-assembly of dna: Some theory and experiments. *Proceedings of the Second DIMACS Workshop on DNA Based Computers*, 1996.

[22] M.A. Zenkova and G.G. Karpova. Imperfectly matched nucleic acid complexes and their biochemical manifestation. *Uspekhi Khimii (PNAS Russia)*, 62(4):414–435, 1993.