Towards a DNA Solution to the Shortest Common Superstring Problem

Greg Gloor[1], Lila Kari[2], Michelle Gaasenbeek[1], Sheng Yu[2]
[1]Department of Biochemistry, University of Western Ontario,
London, Ontario, N6A 5C1 Canada,
ggloor@julian.uwo.ca, mgaasenb@julian.uwo.ca
[2] Department of Computer Science, University of Western Ontario
London, Ontario, N6A 5B7 Canada
lila@csd.uwo.ca, http://www.csd.uwo.ca/~lila, syu@csd.uwo.ca

Abstract

This paper proposes a DNA algorithm for solving an NP-complete problem (The Shortest Common Superstring Problem) by manipulation of biomolecules, and presents partial results of the experiment that implements our algorithm. We also discuss practical constraints that have to be taken into account when implementing the algorithm, propose a coding system as a solution to these practical restrictions, and discuss the control experiments performed for establishing the parameters controlling the specificity of the assay.

1 Introduction

Molecular computing, known also under the name of biomolecular computing, biocomputing or DNA computing, is a new computation paradigm that employs (bio)molecule manipulation to solve computational problems. The excitement generated by the first successful experiment (Adleman 1994, [1]) was due to the fact that computing with biomolecules (mainly DNA) offered an entirely new way of performing and looking at computations: the main idea was that data could be encoded in DNA strands, and molecular biology techniques could be used to execute computational operations. Besides the novelty of the approach, molecular computing has the potential to outperform electronic computers. For example, DNA computing has the potential to provide huge memories: DNA in weak solution in one liter of water can encode 10^{19} bytes, and one can perform massively parallel associative searches on these memories, [7], [20]. Computing with DNA also has the potential to supply massive computational power. A general proposed use of molecular computing is to construct parallel machines where each processor's state is encoded by a DNA strand. DNA in weak solution in one liter of water can encode the state of 10^{18} processors. Moreover, one can perform massively parallel computations by executing recombinant bio-operations that act on all the DNA molecules at the same time. These recombinant bio-operations may be used to execute massively parallel memory read/write, logical operations and also further basic operations, such as parallel arithmetic. Since certain recombinant bio-operations can take minutes to perform, the overall potential for molecular computation is about 1,000 tera-ops, [20].

Despite the progress obtained, substantial obstacles remain before molecular computing becomes an effective computational paradigm. The field is therefore still in the incipient stage of *(i)* testing the suitability of certain molecular biology techniques for computational purposes, and *(ii)* finding a suitable formal model for DNA computing. The research in the field has had therefore, from the beginning, both experimental and theoretical aspects. (For surveys and summaries of the field see [14] and its references, [20], [11], [21], [24].)

The present paper falls into the first category by presenting a DNA algoritm for an NP-complete problem, The Shortest Common Superstring Problem, together with partial results of the experiment that implements the algorithm.

Section 2 introduces The Shortest Common Superstring Problem which has as input a finite set of strings of letters (encoded in DNA sequences) and an integer K. The output is "yes" if there exists a superstring (DNA sequence) of length at most K that contains all the input strings as subsequences, and "no" otherwise.

Section 3 presents our proposed DNA algorithm that solves the problem. The algorithm starts by generating a pool containing all the possible candidates for the superstring role, i.e., all the possible sequences

of length at most K. We proceed then to use a "sieve" strategy, by performing a series of successive steps based on a lab technique called affinity purification. The aim of each step is to retain, from the pool of strands obtained at the previous step, only those that contain a certain input string as a subpattern. The number of such steps will thus be equal to the number of the input strings.

If at the end of this process there is any strand remaining, that strand satisfies all the required conditions and therefore is the sought-after superstring. In this case, the answer to our problems is "yes". If, on the other hand, after performing these steps the solution contains none of the superstring candidates, the answer to our problem is "no".

The solution to this and other problems that use hybridization requires strict attention to the practical problems of hybridization. These practical considerations are addressed in Section 4. The most important aspects are the decreasing specificity of hybridization with increasing oligonucleotide length and the impact of base sequence on the stability of DNA molecules. These limitations were taken into account in our coding of the DNA strands for the input patterns. A second problem is the requirement for a controlled experimental system, in which spurious results can be ruled out. We discuss our experimental assay and show experiments indicating that the final results can be trusted both if the answer is yes and if it is no.

2 The Shortest Common Superstring Problem

The problem chosen for our experiments is *The Shortest Common Superstring Problem*. There were several reasons for our choice. First, the problem is NP-complete, i.e, it is a *hard* computational problem. This means, in particular, that such a problem scales up exponentially and consequently large instances cannot be solved in real-time by electronic computers. (In fact, the question whether real-time, i.e., polynomial time algorithms exist for NP-complete problems is still open.) Finding efficient DNA algorithms for solving it would thus indicate that DNA computing could be quantitatively superior to electronic computing, [11]. Second, the experiment proposed for solving the problem uses readily available reagents and techniques. Last but not least, the problem is a good testing ground for Adleman's bio-operations.

Before formally stating the problem, we summarize the notations used. For a set Σ, card(Σ) denotes its cardinality, that is, the number of elements in Σ. An *alphabet* is a finite nonempty set. Its elements are called *letters* or *symbols*. The letters will be usually denoted by the first letters of the alphabet, with or without indices, i.e., a, b, C, D, a_i, b_j, etc. If $\Sigma = \{a_1, a_2, \ldots, a_n\}$ is an alphabet, then any sequence $w = a_{i_1} a_{i_2} \ldots a_{i_k}$, $k \geq 0$, $a_{i_j} \in \Sigma$, $1 \leq j \leq k$ is called a *string* (*word*) over Σ. The length of the word w is denoted by $|w|$ and, by definition, equals k. The words over Σ will usually be denoted by the last letters of the alphabet, with or without indices, for example x, y, w_j, u_i, etc. The set of all words consisting of letters from Σ will be denoted by Σ^*. For further formal language theory notions and notations the reader is referred to [22].

The Shortest Common Superstring Problem [10]

Input: Alphabet Σ, finite set $R = \{x_1, x_2, \ldots, x_n\}$, $n \geq 1$, of strings from Σ^*, and a positive integer K.
Question: Is there a string $w \in \Sigma^*$ with length $|w| \leq K$ such that each string $x_i \in R$, $1 \leq i \leq n$, is a substring of w, i.e., for all i, $1 \leq i \leq n$, $w = u_i x_i v_i$, where u_i, v_i, are strings in Σ^*?
Comments: The problem remains NP-complete even if card(Σ) $= 2$ or if all $x_i \in R$ have lengths $|x_i| \leq 8$ and contain no repeated symbols. On the other hand, the problem is solvable in polynomial time if all $x_i \in R$ have $|x_i| \leq 2$.

In order to be able to state the problem in molecular biology terms and give it a DNA-based solution, we need a brief introduction of some basic molecular biology notions. For further details of molecular biology terminology, the reader is referred to [16].

DNA (deoxyribonucleic acid) is found in every cellular organism as the storage medium for genetic information. It is composed of units called nucleotides, distinguished by the chemical group, or base, attached to them. The four bases are *adenine, guanine, cytosine* and *thymine*, abbreviated as A, G, C, and T. (The names of the bases are also commonly used to refer to the nucleotides that contain them.) Single nucleotides are linked together end–to–end to form DNA strands. A short single-stranded polynucleotide chain, usually less than 30 nucleotides long, is called an *oligonucleotide* (or, shortly, oligo). The DNA sequence has a *polarity*: a sequence of DNA is distinct from its reverse. The two distinct ends of a DNA sequence are known under the name of the $5'$ end and the $3'$ end, respectively. Taken as pairs, the nucleotides A and T and the nucleotides C and G are said to be *complementary*. Two complementary single-stranded DNA

sequences with opposite polarity will join together to form a double helix in a process called *base-pairing* or *annealing*. The reverse process – a double helix coming apart to yield its two constituent single strands – is called *melting*.

A single strand of DNA can be likened to a string consisting of a combination of four different symbols, A, G, C, T. Mathematically, this means we have at our disposal a 4-letter alphabet $\Sigma = \{A, G, C, T\}$ to encode information. As concerning the operations that can be performed on DNA strands, the existing models of DNA computation are based on various combinations of the following primitive *bio-operations*, [14], [15]:

– *Synthesizing* a desired polynomial-length strand.
– *Mixing:* pour the contents of two test-tubes into a third.
– *Annealing (hybridization):* bond together two single-stranded complementary DNA sequences by cooling the solution.
– *Melting (denaturation):* break apart a double-stranded DNA into its single-stranded components by heating the solution.
– *Amplifying (copying):* make copies of DNA strands by using the Polymerase Chain Reaction (PCR), [8].
– *Separating* the strands by length using a technique called gel electrophoresis.
– *Extracting* those strands that contain a given pattern as a substring by using affinity purification.
– *Cutting* DNA double-strands at specific sites by using commercially available restriction enzymes.
– *Ligating:* paste DNA strands with compatible sticky ends by using DNA ligases.
– *Substituting:* substitute, insert or delete DNA sequences by using PCR site-specific oligonucleotide mutagenesis.
– *Detecting and Reading* a DNA sequence from a solution.

We are now ready to formulate the Shortest Common Superstring Problem in molecular biology terms:

The Shortest Common Superstring Problem in Molecular Terms
Given n oligonucleotide strings $x_1, x_2, \ldots, x_n$ of arbitrary lengths, and a positive number K, is there a nucleotide sequence w of length at most K that contains all the oligonucleotides $x_1, x_2, \ldots, x_n$ as subsequences?

The solution to this problem also provides a method for finding the minimum-length sequence containing all the given oligonucleotides. Note that such a sequence always exists: the catenation $x_1 x_2 \ldots x_n$ of all nucleotides strings is a nucleotide sequence containing all the given oligonucleotides. Due to possible overlaps, this catenation is not necessarily the minimal (shortest) sequence that contains all given oligonucleotides. The minimal sequence is called the *shortest common superstring* of the given oligonucleotides.

Example: If $\Sigma = \{G, T\}$, $R = \{GTG, TGT, GTT\}$ and $K = 5$, the answer to the problem is "yes". Indeed, the superstring $GTGTT$, of length 5, contains all the input strings as subsequences.

On the other hand, if we let Σ and R as in the previous example but change the input value of K to 4, the answer to the problem becomes "no". Indeed, no string of length 4 can be found that contains all the input strings as subpatterns.

3 Biomolecular solution

This section contains a DNA algorithm which we developed to solve The Shortest Common Superstring Problem, and partial steps of its practical implementation. The experiment uses readily available reagents and techniques. Compared to standard protocols, the main difference is the number of reactions conducted entirely in vitro prior to cloning of the products. Careful optimization of each step will be required to ensure maximum specificity. However, these reactions are possible to perform entirely within the established parameters of each enzyme.

DNA algorithm:

Step 1. Encode all the strings $\{x_1, x_2, \ldots, x_n\}$ of the set R in DNA strands.

Step 2. Generate all the possible DNA strands w of length between $max\ \{|x_i|, 1 \leq i \leq n\}$ and K.

Step 3. Let x_1 be a string of R. From the string population of candidates generated in *Step 2* select only those strands that contain x_1 as a subsequence. From the newly obtained string population, select only those strings that contain $x_2 \in R$ as a subsequence, etc. Repeat the step for each strand x_i in R, $1 \leq i \leq n$.

Step 4. If, after *Step 3*, there is any strand w remaining (which means that w contains all $x_i \in R$, $1 \leq i \leq n$, as subsequences), say YES, otherwise NO.

Implementation:

Step 1. Encode the strings x_i, $1 \leq i \leq n$, as DNA fragments using A, C and T residues only. Each of the synthesized strings is biotinylated (has a biotin tag attached to it) at a single site to permit recovery of the oligonucleotide and the bound complementary sequence from solution, [17], [23].

Step 2. Generate all the possible length k nucleotide sequences, starting with $k = max\{|x_i|, 1 \leq i \leq n\}$ using T, G and A residues only: each of them is a candidate for the complement of the common superstring. The structure of each of the complements of candidates will be: a marker sequence α (20 nucleotides long) followed by a length k nucleotide sequence and ending with another marker sequence β (20 nucleotides long). The marker sequences α and β contain all the bases A, C, G, and T and are carefully chosen so that overlaps with each of the strings x_i are avoided.

If k were 25, the total number of different sequences represented in this step would be 3^{25}. Put another way, 1.4 *pMole* of oligonucleotide would be required to encode every sequence uniquely. In practice, we use approximately 40 *pMoles* of oligonucleotide to ensure that each sequence is represented more than 20 times. The initial hybridization can thus occur in approximately 40 μl of solution with each oligo at a concentration of 1 $\mu M/l$. The exponential increase in sequence complexity dictates an obvious practical upper limit on this approach, as for $k = 30$ we would need 10 ml of solution while for $k = 35$ we would need approximately 2.5 l.

The procedure in *Step 2* is repeated for increasing values of k, i.e., $k+1, k+2 \ldots$ up to K. At the end of this procedure we obtain a pool containing all the complements of possible superstring candidates, flanked by the marker sequences α and β.

Step 3. Test the existence of a segment of DNA of length at most K, that contains the complements of all the sequences representing the n given strings.
This will be accomplished by sequential selection from the pool of randomly generated sequences obtained in *Step 2*, by hybridization (annealing) with the individual given oligonucleotides:
3a) Denature: Mix the oligo that represents string x_1 with the pool of strands generated at *Step 2* under denaturing conditions at 92°C, [19], [27]. The purpose of this step is to prevent intra-chain annealing and to have the DNA sequences fully single stranded and available for pairing.
3b) Anneal: Bring the mixture to a temperature that is about 5°C below the Tm (melting temperature, i.e., the temperature at which the two DNA strands of double-stranded DNA separate) of the oligo representing x_1. The Tm can be determined empirically, or can be calculated for a given oligonucleotide and salt concentration, [19], [27]. As a result of this step, the sequence representing x_1 will anneal to all strands that contain the complement of x_1 as a subsequence.
3c) Select: Mix the annealed mixture with an avidin conjugated to a solid support, to retain all the biotinylated oligonucleotides and the bound complements of superstring candidates. The avidin, fixed to the solid support, will bind every biotin tag attached to the oligo representing x_1. The oligo representing x_1 is, in turn, annealed to all strands containing its complement as a subsequence. Consequently, all the oligonucleotides containing the complement of x_1 will be retained.
3d) Wash: wash away all unbound oligonucleotides, as they fail to contain the complement of x_1 as a subsequence.
3e) Remove: Raise the temperature to greater than the Tm of the oligonucleotide representing x_1 and collect the eluent. Raising the temperature will induce melting of the bonds between the oligo representing x_1 and the strands annealed to it. The eluent will contain thus all the oligos containing the complement of x_1 as a subsequence.
3f) Repeat: Repeat the steps *3a)* – *3e)* for the oligos representing the strings $x_2, x_3, \ldots, x_n$. Each sucessive eluent will contain subsequences complement to all of the so-far tested oligonucleotides.

Step 4. After the sequential selection steps are completed, the final product will be amplified by PCR, cloned and sequenced to find the answer:
4a) Amplify: The eluent from repeat n in *Step 3* will have only a few molecules of the desired sequences. This will be amplified with the poymerase chain reaction using oligonucleotide primers that bind to the unique sequences α and β that flank the complements of candidate sequences.
4b) Sequence: The amplified products will be cloned and sequenced to confirm the length of the sequence, and to confirm that the sequence does contain the complements of the given strings. Alternatively, the products could be separated by size on a polyacrylamide gel and sequenced directly. The complement of the obtained string, if exists, represents the sought-after common superstring.

If necessary, the eluents after each *Step 3e)* can be amplified with the PCR to progressively increase the amount of DNA sequences for the next cycle of reactions. Asymmetric PCR can be used to generate the required single strands [12].

One possible problem is the error rate of hybridization and of DNA polymerization in the PCR, [9], [12], [13], [26]. Should the error rate resulting from Taq polymerase be a problem, this could be addressed, in part, by using a thermostable DNA polymerase with a low error rate such as Pfu polymerase, [18], [26]. We can get the correct answer by sequencing a number of the clones, to identify families of DNA sequences and to accept the consensus sequence as the answer. This method of averaging corrects the random errors. A consensus sequence could also be generated by direct sequencing of the PCR products.

Pilot experiments:

We are in the process of testing two chosen candidate oligonucleotide sequences, derived by inspection, one of them containing the complements of all the given sequences and another lacking the complement of at least one sequence. The test sequences are flanked by the marker sequences α and β. These experiments will ensure that the procedure can reliably produce a positive and a negative result, both of which are significant.

4 Practical considerations

The *stability* of a double-stranded oligonucleotide sequence depends on *(i)* the number of hydrogen bonds between base pairs (two for A binding to T, and three for G binding to C), and *(ii)* the attraction forces that exist between adjacent bases (the adjacent pair GC is the strongest, while the adjacent pair TA is the weakest), [19]. Thus every additional base-pair increases the stabilization energy by a predictable amount, [19].

If a double-stranded oligonucleotide has a mismatch, some or all of the stabilization energy that would have been present in the case of the perfect match is lost. Unfortunately, the effect of a single base mismatch on the stability of a short oligonucleotide cannot be predicted at present for all mismatches. Effects of individual mismatches range from negligible to extreme for the same mismatch. The nucleotides flanking the mismatched base-pair dramatically affect stability [2, 3, 4, 5, 6].

The *specificity* of an oligonucleotide is defined as the difference between the annealing temperature of the oligonucleotide with its perfectly matched complement, and the annealing temperature of the same oligo with a "near"-complement with which it has a one-base mismatch. Note that, the greater the stability of an oligonucleotide, the smaller its specificity, as stability means greater tolerance for mismatches.

The main bio-operation used in our experiment is annealing. The thermodynamic parameters controlling DNA-DNA annealing, briefly introduced above, are known in sufficient detail to allow reasonably precise predictions of DNA-DNA annealing reactions, [19]. These thermodynamic parameters show that there are constraints which are unavoidable for DNA strands that contain the 4 naturally occurring bases A, C, G, T. (Note that Wetmur, [27], provides a formula for converting the thermodynamic information into annealing/melting temperatures.) There are three main constraints on the use of oligonucleotides for annealing, which we have to keep in mind when desigining our oligonucleotide strings.

The first practical requirement is that the annealing temperatures be in a fairly narrow range of temperatures. Biological molecules are rapidly destroyed at high temperatures. For example, DNA samples heated to 95 degrees for extended periods ($>$ 5 minutes) become poor templates for PCR amplification. In addition, virtually all proteins are irreversibly inactivated when exposed to temperatures higher than 42 degrees. On the other hand, temperatures must be high enough for the biological reactions required for molecular computing to occur in seconds and minutes rather than hours and days. In practice this means working in a temperature range between 10 and 75 degrees at the extremes, although almost all biological reactions except PCR work best between 30 and 40 degrees.

The second practical requirement is that similar length oligos should have similar annealing temperatures within the optimal temperature range. This allows dealing with same-length oligos at similar annealing temperatures, which in turn greatly simplifies the experimental procedures.

The annealing properties of an oligonucleotide vary with the length, composition (proportion of each base, sequence (arrangement of bases) of the oligonu-

cleotide, Na^+ concentration ([Na^+]) and oligonucleotide concentration [19], [27]. The Na^+ concentration can be chosen by the investigator to meet the needs of the experiment. Higher [Na^+] results in increased annealing temperatures. Likewise, the oligo concentration can be optimized by the investigator. If the Na^+ or oligo concentration becomes too high, then specificity is lost. In practice, the Na^+ concentration is usually held below 1 M/l and the oligo concentration is near 1 $\mu M/l$.

Two oligonucleotides with an identical base composition but different arrangements of the bases will give widely varying annealing temperatures as shown in *Table 1*.

Table 1: Variation in DNA-DNA annealing/melting temperature with base sequence, [25].

Oligonucleotide	Tm
TTTTTAAAAA	18 degrees
AAAAATTTTT	20 degrees
ATATATATAT	8 degrees
TATATATATA	7 degrees

[C]= $1\mu M/l$, [Na^+]= $0.2M/l$.

As can be seen from *Table 1*, this variation is caused by the non-random arrangement of bases in short oligonucleotides. Therefore, one way to reduce this variation is to encode the information in DNA strings of sufficient length that the sequences in the string are essentially random. A second solution is to restrict the encoding such that all arrangements of the bases give similar annealing temperatures.

The third requirement is that the specificity of the oligonucleotides used be maximized. This induces another constraint, as the specificity of an oligonucleotide for its exact complement decreases with increasing oligonucleotide length. In other words, the longer the oligonucleotide, the greater the chance that it will bind not only to its perfect complement, but to a sequence closely resembling its complement. In fact, the number of tolerated mismatches increases with length.

In addition, the position of the mismatched base pair in the oligonucleotide has a strong effect on the stability of an oligonucleotide duplex.

As an example, we can calculate the annealing temperatures of oligonucleotides composed of only A residues of varying lengths. The results of such a calculation, based on [25], is shown in *Table 2*:

Table 2

Oligo	Tm	Diff.
AAAAAAA	8	
AAAAAAAA	17	9
AAAAAAAAA	25	8
AAAAAAAAAA	31	6
AAAAAAAAAAA	35	4
AAAAAAAAAAAA	39	4
AAAAAAAAAAAAA	43	4
AAAAAAAAAAAAAA	45	2

Oligo concentration = 1 $\mu M/L$, [Na^+] = 1.0 M/L.

As can be seen in *Table 2*, the difference in Tm between an oligonucleotide of length N and length $N+1$ decreases with increasing oligonucleotide length. This simple illustration shows two properties of oligonucleotide annealing. Firstly, there is a strong dependence of Tm on oligonucleotide length. Secondly, each additional base contributes sucessively less to the overall Tm. As an approximation, an oligonucleotide with a single base mismatch at either end has the Tm of an identical oligonucleotide that is missing the terminal mismatched base. This simple case shows how the specificity of an oligonucleotide for its exact complement decreases with increasing length.

Note however, that mismatches at the end of the oligonucleotide are the least destablilizing. Allawi and SantaLucia ([3, 4, 5, 6]) examined the stability of oligonucleotides mismatched in the middle of the sequence. *Table 3* shows a sample of the mismatched oligonucleotides tested, their Tm and the Tm of the perfectly complementary sequence.

Table 3

Oligonucleotide	Tm (if mismatch.)	Tm (if compl.)
GCTCgCAGG t	52	67
CATGAtGCTAC c	47	59
GTAGTcACATG a	46	62
GGAGgCACG a	56	66

In this case the difference in Tm is always at least 10 degrees, which is a much greater difference than for an oligo of length $N-1$.

These results show that it will be difficult to determine the specificity of an oligonucleotide to a mismatch at an arbitrary site. In practice, specificities of less than 5 degrees are essentially useless, and we are aiming at specificities of at least 10 degrees.The phenomenon can be generalized to heteropolymeric oligonucleotides: longer oligonucleotides can hybridize and be relatively stable even with several base mismatches between the oligo and its near-complement.

One solution to the third requirement of maximizing specificity is thus to use short oligonucleotides. Note that this solution contradicts the first requirement and one of the solutions to the second requirement. There is a practical trade-off between melting temperatures and specificity, and arbitrary mismatches cannot be distinguished by annealing once the oligo becomes longer than about 10 bases.

Coding for an annealing reaction

Being aware of these three constraints we attempted to choose oligonucleotides with optimized coding to give both a small temperature variation within the optimal range for the annealing temperatures, and maximum specificity for our annealing reactions.

To do this, we used two strategies. First, we chose strings of length 9. This was long enough to give biologically meaningful annealing temperatures, satisfying thus the first requirement. To deal with the second requirement, we chose the alternative solution to using long sequences (which would have decreased specificity). By using the three bases and avoiding the most stable (GC) and most unstable (TA) sequences, we ensured that all stabilization energies were located in a small range, which implied that the annealing/melting temperatures were within a small interval. As we avoided the GC and TA sequences, the fact that annealing temperatures were located within a small range greatly eased the requirement for randomness and thus for long sequences.

The third requirement was taken care of by choosing oligonucleotides that were short enough to ensure a good specificity.
The encodings for our initial substrings are:

$$\begin{aligned} x_1 &= 5'-CATCATCAT-3' \quad Tm = 26 \\ x_2 &= 5'-AAATTTCAT-3' \quad Tm = 20 \\ x_3 &= 5'-CCTTCAAAA-3' \quad Tm = 27 \end{aligned}$$

$[\mathrm{Na}^+] = 0.2M/l$, oligo concentration $= 1\mu M/l$

The specificity for x_1 (respectively for x_2 and x_3) is 5 degrees (respectively 5 degrees and 6 degrees) if the mismatch is at the 3' end, and 10 degrees (respectively 7 degrees and 10 degrees) if the mismatch is at the 5' end. If the mismatch is somewhere in middle of the oligonucleotide, the specificity is usually greater than 10 degrees. Thus, our encoding systems provides a good compromise between stability and specificity.

The need for a controlled experiment

A major concern in any biologically-based experiment are false positive and false negative results. Therefore, the experimentor must design the experiment so that all conceivable causes of false results can be identified and accounted for when interpreting the experiment. In biology these parts of the experiment are called *controls* and serve as internal quality control stardards by which the reliability of the result can be determined.

A PCR assay for recovering annealing products

Two different (complements of) superstring controls were constructed that contained the same base composition, but that differed in base sequence.

$$s_1 = 5'-\alpha TTTGATGATGAAATTTTGAAGG\beta-3'$$

$$s_2 = 5'-\alpha TGTGGTGATGATATTGTGAAGG\beta-3'$$

where

$$\alpha = 5'-GCCGAAGCTTACCGAAGTAT-3'$$

$$\beta = 5'-GCACTTATTGCAAGCATACG-3'$$

One of the chosen strings, s_1, is known to contain the complements of all the substrings x_1, x_2, x_3, while s_2 has at least one mismatch with each of the substrings. The superstrings are flanked by two unique sequences α and β which provide tags for PCR amplification following the selection. We are in the process of proving that the annealing reaction between the substrings and the perfectly matched superstring s_1 always gives a PCR product, and that the annealing reaction between the substrings and the imperfectly matched superstring s_2 never gives a PCR product. We will attempt the full experiment once we have established the annealing parameters controlling the specificity of the assay.

It is essential that any superstrings that are bound by the substrings be identified. We have established a PCR-based assay for the superstrings. This assay can

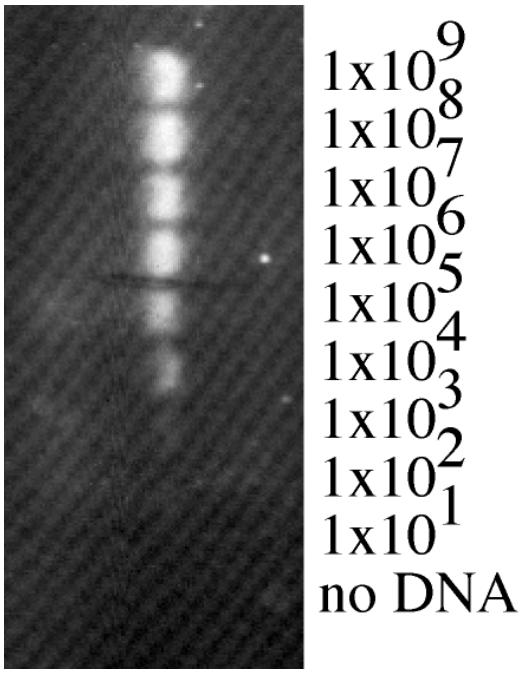

Figure 1: Detection of superstring s_1 by a PCR-based assay. PCR reactions were set up with decreasing numbers of s_1 as the template. Amplification was carried out for 25 cycles with a cycle profile of 95 degrees for 45 seconds, 55 degrees for 60 seconds, 72 degrees for 30 seconds. The products were run on a 4% NuSieve agarose gel and stained with ethidium bromide.

easily detect 10000 molecules of superstring following 25 cycles of PCR (Figure 1). When the PCR contained 1000 molecules, the signal was very faint, and does not show up well in the reproduction. No signal was detected with less than 1000 molecules in the reaction, or when DNA was not added to the PCR reaction.

References

[1] L.Adleman. Molecular computation of solutions to combinatorial problems. *Science* v.266, Nov.1994, 1021–1024.

[2] H. Allawi et al. Thermodynamics of Internal G:T mismatches. *Biochemistry* 36(1997), 10581-10594.

[3] H. Allawi, J. SantaLucia. Thermodynamics of internal C-T mismatches in DNA. *Nucleic Acids Res.* 26(1998), 2694-2701.

[4] H. Allawi, J. SantaLucia. Nearest-neighbor thermodynamic parameters for internal G-A mismatches in DNA. *Biochemistry* 37(1998) 2170-2179.

[5] H. Allawi, J. SantaLucia. Nearest-neighbor thermodynamic parameters for internal A-C mismatches in DNA: sequence dependence and pH effects. *Biochemistry* 37(1998), 9435-9444.

[6] H. Allawi, J. SantaLucia. Thermodynamics of internal G-T mismatches in DNA. *Biochemistry* 36(1997) 10581-10594.

[7] E.Baum. Building an associative memory vastly larger than the brain. *Science*, vol.268, April 1995, 583–585.

[8] C.W.Dieffenbach, G.S.Dveksler, Eds. *PCR primer: a laboratory manual*, Cold Spring Harbor Laboratory Press, 1995, 581-621.

[9] H.A.Erlich. *PCR Technology*, Stockton Press, New York, 1989.

[10] M.Garey, D.Johnson. *Computers and intractability: a guide to the theory of NP-completeness.* Freeman and Co., San Francisco, 1979.

[11] D.K.Gifford. On the path to computation with DNA. *Science* 266(Nov.1994), 993–994.

[12] U.Gyllensten and H.A.Erlich. 1988, Generation of single-stranded DNA by the polymerase chain reaction and its application to direct sequencing of the HLA-DQA locus. *Proc.Nat.Acad.Sci. USA*, 85(1988), 7652-7656.

[13] M.A.Innis, et al. *PCR Protocols: A Guide to Methods and Applications,* Academic Press, San Diego, 1990.

[14] L.Kari. DNA computing: arrival of biological mathematics. *The Mathematical Intelligencer*, vol.19, nr.2, Spring 1997, 9–22.

[15] L.Kari. From Micro-Soft to Bio-Soft: Computing with DNA. *Proceedings of BCEC'97 (Bio-Computing and Emergent Computation)* Skovde, Sweden, World Scientific Publishing Co., 146–164.

[16] J.Kendrew et al., eds. *The Encyclopedia of Molecular Biology*, Blackwell Science, Oxford, 1994.

[17] T.Maniatis, et al. *Molecular Cloning: A laboratory manual*, Cold Spring Harbor Press, New York, 1982.

[18] P.Mattila, et al., 1991, Fidelity of DNA synthesis by the Thermococcus litoralis DNA polymerase–an extremely heat stable enzyme with proofreading activity. *Nucleic Acids Research*, 19(1991), 4967–73.

[19] Owczarzy et al. Predicting Sequence-Dependent Melting Stability of Short Duplex DNA Oligomers. *Biopolymers*, 44(1998), 217 -239.

[20] J. H. Reif, Paradigms for Biomolecular Computation. *Unconventional Models of Computation*, C.S.Calude, J.Casti, M.J.Dinneen, Eds., Springer Publishers, 1998, 72–93.

[21] H.Rubin. Looking for the DNA killer app. *Nature*, 3(1996), 656–658.

[22] A.Salomaa. *Formal Languages*. Academic Press, New York, 1973.

[23] J.Sambrook, et al. *Molecular Cloning: A laboratory manual, second edition*, 1989.

[24] W.Smith. DNA computers in vitro and in vivo, *1st DIMACS workshop on DNA based computers*, Princeton, 1995. In *DIMACS series*, vol.27 (1996), 121–185.

[25] J. SantaLucia. A unified view of polymer, dumbbell and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc. Nat'l Acad. Sci. USA*, 95(1998), 1460-1465.

[26] K.R.Tindall and T.A.Kunkel. Fidelity of DNA synthesis by the Thermus aquaticus DNA polymerase. *Biochemistry* 27(1988), 6008–13.

[27] J.G.Wetmur. DNA Probes: Applications of the principles of nucleic acid hybridization. *Crit.Rev.Biochem. and Mol.biol.*, 26(1991), 227–259.