The fidelity of template-directed oligonucleotide ligation and its relevance to DNA computation

Kenneth D. James, Amy R. Boles, Dorothy Henckel and Andrew D. Ellington[1],*

Department of Chemistry, Indiana University, Bloomington, IN 47405, USA and [1]Department of Chemistry, Institute for Cellular and Molecular Biology, University of Texas at Austin, 2500 Speedway, Austin, TX 78712, USA

Received June 23, 1998; Revised August 5, 1998; Accepted September 28, 1998

ABSTRACT

Several different computational problems have been solved using DNA as a medium. However, the DNA computations that have so far been carried out have examined a relatively small number of possible sequence solutions in order to find correct sequence solutions. We have encoded a search algorithm in DNA that required the evaluation of >16 000 000 possible sequence solutions in order to find a single, correct sequence solution. Experimental evaluation of the search algorithm revealed bounds for the accuracies of answers to other large, computationally complex problems and suggested methods for the optimization of DNA computations in general. Short oligonucleotide substrates performed substantially better than longer substrates. Large, computationally complex problems whose evaluation requires hybridization and ligation can likely best be encoded and evaluated using short oligonucleotides at mesophilic temperatures.

INTRODUCTION

Following the demonstration by Adleman (1) that a directed Hamiltonian path problem (HPP) could be encoded in DNA and evaluated, the use of nucleic acids for computational purposes has been the focus of extensive speculation. DNA has been touted as an appropriate medium for computational problems ranging from the simple addition of integers (2) to several different classes of non-deterministic polynomial (NP) time complete problems (3–5). The DNA computations that have been carried out have frequently relied on the molecular biology operations of hybridization, ligation and PCR amplification to generate solutions. For example, the HPP examined by Adleman (1) involved an encoded graph in which seven DNA oligonucleotides represented seven vertices or ‘cities’ and 13 complementary, partially overlapping DNA oligonucleotides represented unidirectional edges or ‘paths’ between the cities. In order to determine if a Hamiltonian path (in which each of the seven cities appeared only once) was present on the encoded graph, the oligonucleotide cities were juxtaposed with one another by complementary oligonucleotide paths and then covalently joined by enzymatic ligation (Fig. 1a). The binary joinings were further catenated to form longer paths. The existence of a Hamiltonian path on the encoded graph was confirmed by size separation and hybridization analysis of the ligation products.

However, to date no large, computationally complex problem has yet been encoded in DNA and solved by molecular biology methods (6). The DNA computations that have been carried out have relied on a relatively small number of oligonucleotides that represent a correspondingly small number of either possible or correct sequence solutions. The HPP was encoded with 20 oligonucleotide strings, a Maximal Clique problem was encoded in 25 strings (4), a binary addition was carried out with seven strings (2) and a method for encoding DNA words was tested with 108 strings (7). In consequence, it is unclear whether molecular biology methods such as hybridization, ligation and amplification will still yield accurate answers when scaled to larger, computationally complex problems. For example, the HPP that has been solved involved only 637 possible joinings of oligonucleotide cities (13 paths × 7 cities × 7 cities, assuming that oligonucleotide paths were distinguished from oligonucleotide cities and that there was no shift in the register of base pairing). In contrast, a similar HPP with 100 (rather than seven) cities would have required the simultaneous evaluation of up to 99 000 000 possible joinings of oligonucleotide cities (9900 paths × 100 cities × 100 cities). Depending on how many Hamiltonian paths originally existed on the encoded graph, the unprogrammed, fortuitous joining of two cities could drastically decrease the reliability of extracted solution sets. In a worst case scenario, if no Hamiltonian paths existed on an encoded graph, then the unprogrammed joining of two cities could potentially produce a false Hamiltonian path and thus give a categorically wrong answer to the HPP. Adleman (1) also notes that so-called ‘pseudopaths’ might form and would detract from the accuracy of DNA computations.

To determine if large, computationally complex problems might eventually be attempted using DNA as a medium, we devised a search algorithm in which literally millions of possible solutions were simultaneously appraised. This search algorithm again relied on the molecular biology operations of hybridization, ligation and PCR amplification. A query sequence (template) was mixed with an extremely diverse oligonucleotide pool and members of the pool that were solutions to the query were ligated in place and subsequently amplified. Reaction conditions have been identified that yielded almost exclusively accurate sequence solutions to the search algorithm, while departure from these conditions yielded a high proportion of inaccurate sequence solutions.

*To whom correspondence should be addressed. Tel: +1 512 471 6445; Fax: +1 512 471 7014; Email: andy.ellington@mail.utexas.edu

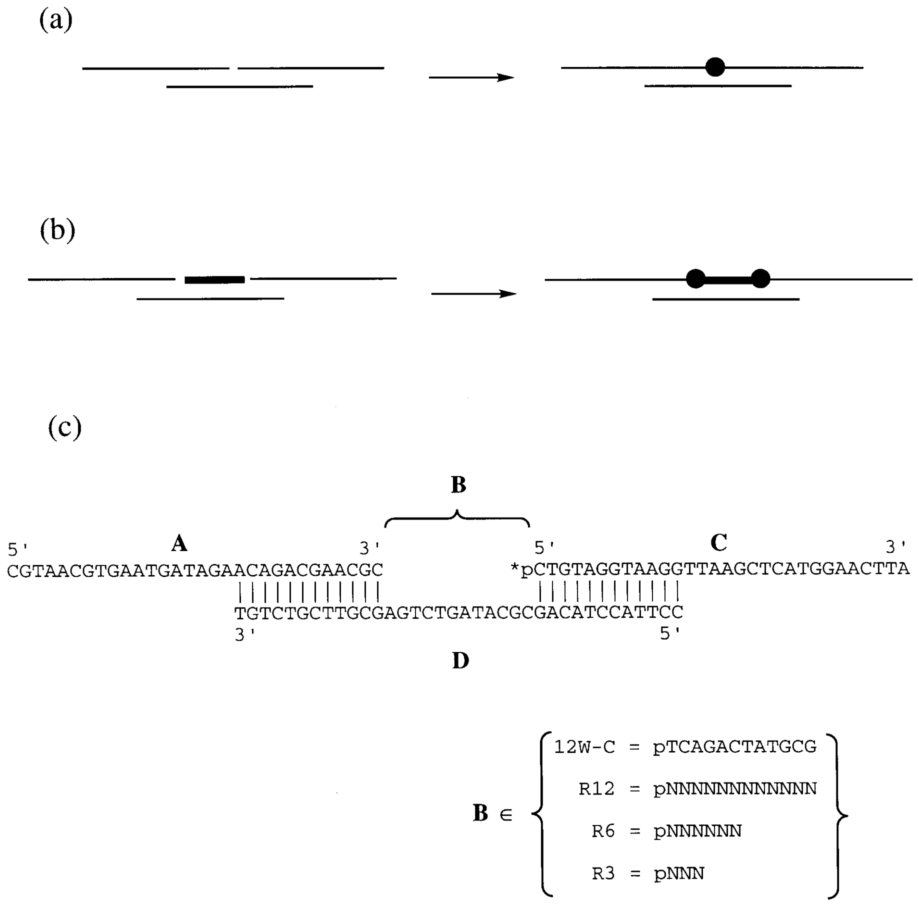
(a)
(b)
(c)
B
5'
A
3'
5'
C
3'
CGTAACGTGAATGATAGAACAGACGAACGC
*pCTGTAGGTAAGGTTAAGCTCATGGAACTTA
TGTCTGCTTGCGAGTCTGATACGCGACATCCATTCC
3'
5'
D
B ∈
12W-C = pTCAGACTATGCG
R12 = pNNNNNNNNNNNN
R6 = pNNNNNN
R3 = pNNN

Figure 1. Schemes for DNA computation. (**a**) The basic molecular biology operations required for the solution of DNA computations such as the Hamiltonian path (1), Maximal Clique and addition problems. Two oligonucleotides are juxtaposed by a template and joined to one another by ligation. (**b**) The basic molecular biology operations required for the solution of the search algorithm described in this paper. Three oligonucleotide substrates are juxtaposed by a template and joined to one another by ligation. The terminal oligonucleotide substrates are constant sequences, while the internal oligonucleotide substrate (bold line) is a pool of dodecamers. (**c**) The sequences of the oligonucleotides used for the evaluation of the search algorithm. The precise Watson–Crick complement of the single-stranded region in the gapped duplex is 12W-C. Each position in the random sequence pools R12, R6 and R3 is shown as N, which represents an approximately equimolar mixture of the four canonical deoxyribotides. The 5′-ends of oligonucleotides C, 12W-C, R12, R6 and R3 are phosphorylated (p). In several experiments, the 5′-end of C was radiolabeled (*p).

MATERIALS AND METHODS

Materials and reagents

All oligonucleotides were synthesized on an ABI PCR Mate DNA synthesizer. Reagents were purchased from Glen Research (Sterling, VA); columns were obtained from Cruachem (Dulles, VA). Terminal 5′-phosphates were added to oligonucleotides during automated synthesis. Oligonucleotides longer than 15 residues were purified by gel electrophoresis. T4 polynucleotide kinase (New England Biolabs, Beverly, MA) and [γ-^{32}P]ATP (NEN Research Products, Boston, MA) were used to end-label oligonucleotides. DNA ligase purified from *Escherichia coli* was obtained from New England BioLabs. Taq polymerase was obtained from Promega (Madison, WI). Sequenase was obtained from US Biochemical (Cleveland, OH).

Enzymatic ligations

Constant sequence oligonucleotides were combined in a thin-walled PCR tube at a concentration of 20 µM each in a final volume of 4.5 µl of 50 mM Tris–HCl, pH 7.8, 10 mM $MgCl_2$, 10 mM dithiothreitol, 26 µM NAD^+ and 125 ng bovine serum albumin. If a random pool was used in the reaction, the pool was added in excesses of 5-fold for the dodecamers, 25-fold for the hexamers and 125-fold for the trimers. These values were based on the efficiencies of product formation in trial reactions with perfectly paired dodecamers, hexamers and trimers. The oligonucleotides were denatured at 95°C for 3 min and cooled to room temperature over 15 min in an MJ Research Inc. (Watertown, MA) Minicycler. The solutions were then either kept at 25°C in the Minicycler or cooled to 16°C and DNA ligase (1 U) was added. The reactions were quenched at times ranging from 15 s to 1 h by the addition of 5 µl of denaturing dye (bromophenol blue, 7 M urea).

Amplification of sequence solutions

Ligation products were separated from unligated oligonucleotides by gel electrophoresis (12% polyacrylamide; 19:1 mono-acrylamide:bis-acrylamide) in the presence of 7 M urea and TBE buffer. Radioactive bands of appropriate length were excised from the gels and eluted overnight at 37°C into 0.3 M NaCl. The eluted oligonucleotides were precipitated by the addition of 3 vol of ethanol followed by centrifugation at 4°C. The pellets were washed with 70% ethanol, air dried and resuspended in 5 µl water.

The ligated oligonucleotide products were preferentially amplified using PCR: 2 µl of the isolated ligation product was added to a reaction mix that contained 5% acetamide, 0.05% NP-40, 200 µM dNTPs, 500 nM of each primer (forward, 5′-CGTAACGTGAATGATAGA; reverse, 5′-TAAGTTCCATGAGCTTAA), 1.5 mM $MgCl_2$, 50 mM KCl, 10 mM Tris–HCl, pH 8.3. In each thermal cycle, the temperature was held at 94°C for 30 s, 45°C for 30 s and 72°C for 30 s.

Identifying sequence solutions

Double-stranded PCR products were cloned using a Zero Blunt-End Cloning Kit (Invitrogen, Carlsbad, CA). Plasmids containing inserts were prepared for sequencing using a common protocol (8). The sequences of the inserted oligonucleotides were determined using a cycle sequencing kit (Epicentre Technologies, Madison, WI).

RESULTS AND DISCUSSION

Experimental strategy and conditions

The basic molecular biology operations performed during the evaluation of HPPs or other NP complete problems encoded in DNA have been the template-directed juxtaposition and ligation of oligonucleotides, as shown in Figure 1a, followed by amplification and sieving by hybridization. We have mimicked this DNA computation by encoding a search algorithm in DNA. The evaluation of the encoded search requires the hybridization and ligation of oligonucleotides drawn from a pool of sequences, as shown in Figure 1b, followed by amplification and sequencing. In this case, however, one of the hybridization partners (the template or query sequence) is a single sequence and the other partner (the substrate) is a multitude of different sequences (substrate, up to 4^{12} = 16 777 216 sequences). The solution of the search algorithm requires the evaluation of >32 000 000 possible ligation events (~16 000 000 sequences × 2 ligations/sequence). Therefore, we believe the procedures required for the evaluation of the search algorithm approximate the procedures that would be required for the evaluation of a large, computationally complex problem. The fidelity of evaluation of the search algorithm should provide some estimate of the fidelity of evaluation of other DNA computations.

Operationally, the query sequence was a dodecamer string embedded within a gapped DNA duplex (Fig. 1c). The query sequence contained the four canonical bases in equal measure and thus allowed the assessment of all possible base pairings. A random sequence DNA pool that contained all possible dodecamers served as a potential solution set. The random sequence pool was mixed with the gapped duplex and oligonucleotide substrates that hybridized across from the gap were fixed in place by ligation. While we had previously carried out similar studies using chemical ligation (9), most DNA computations that have been proposed or implemented rely on enzymatic ligation. Therefore, *E.coli* DNA ligase was employed as a coupling reagent. A mesophilic ligase was chosen rather than a thermophilic ligase because the melting temperatures of even perfectly paired dodecamer substrates were much lower than the temperature optima of commercially available thermostable ligases, such as the DNA ligase from *Thermus aquaticus*. In addition, the *E.coli* ligase was chosen over more popular molecular biology enzymes, such as T4 DNA ligase, because it had been shown to inefficiently catalyze blunt-end ligation events (New England Biolabs, Beverly, MA) that could have given rise to artifactual products.

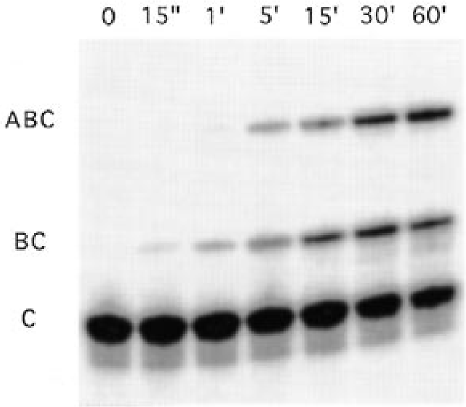

Figure 2. Time course of enzymatic ligation. Oligonucleotides A, 12W-C and *pC (Fig. 1) were mixed with template D at 25°C, a ligation reaction was initiated and reaction products were separated on a denaturing polyacrylamide gel as a function of time. The intermediate ligation product between pB and C (BC) appeared first, followed by the full-length ligation product between A, pB and *pC (ABC). Because of the position of the radiolabel, the intermediate ligation product between A and B could not be observed, but is likely present.

Sequence solutions to the search algorithm were preferentially amplified using PCR. Several steps were taken to avoid amplification artifacts and cross-contamination of amplification reactions. First, primer binding sites were presented within single-stranded overhangs to guard against the inadvertent amplification of template strands. Second, only those oligonucleotides that were programmed to ligate were phosphorylated. Third, full-length ligation products were gel isolated prior to amplification and sequencing. Fourth, control ligation products (Fig. 2) were never present on the same gel as experimental ligation products. Finally, portions of gels adjacent to lanes containing ligation products were also excised and putative ligation products were isolated. When amplification reactions were seeded with these extracts no amplification products were observed.

In order to determine whether our search algorithm could be successfully evaluated, a constant sequence dodecamer that corresponded to the correct sequence solution was first mixed with the gapped duplex containing the query sequence, a ligation reaction was initiated and the products were analyzed by gel electrophoresis. Only the correct full-length product and an intermediate leading to the correct full-length product were observed (Fig. 2). The formation of full-length product was monitored as a function of time. The ligation product could be observed as early as 1 min, but continued to accumulate for >1 h. Based on these experiments, sequence solutions to the search algorithm were extracted at two time points: 1 min and 1 h. By examining both time points we could potentially determine whether there was a trade-off between the efficiency of evaluation of the search algorithm (the amount of product produced) and the accuracy of evaluation of the search algorithm (the proportion of fully complementary products obtained). Since large, computationally complex problems encoded in DNA may require the efficient joining of hundreds to thousands of different oligonucleotides, relatively low reaction temperatures (16 and 25°C) were chosen to maximize the efficiency of ligation (8). The choice of relatively low temperatures was also prompted in part by a desire to compare results with the dodecamer pool with results with hexamer and trimer pools (below); the shorter pools would not have efficiently hybridized to the gapped duplex template at temperatures significantly >25°C (9). The fact that the *E.coli* ligase was active at reduced temperatures again made this enzyme an ideal choice for these experiments. Finally, the search algorithm was performed at the two slightly different temperatures to determine whether its evaluation was robust to reaction conditions or was easily perturbed by subtle alterations in the molecular biology 'hardware'.

Solutions to the search algorithm

The sequences of four sets of amplified ligation products, corresponding to ligation at 16 or 25°C for 1 min or for 1 h, are shown in Figure 3a. As is obvious from even a cursory examination of the sequence data, the fidelity of evaluation of the search algorithm under these conditions was relatively low. These results are further categorized in terms of potential base pairings in Figure 3b. The most frequent mismatches that were observed were G:T wobble pairings, but almost all other mismatches were observed as well. On average, each dodecamer sequence solution to the search algorithm contained 3.3 mistakes (mismatches, deletions or insertions; Table 1). The prevalence of multiple mismatches relative to perfectly paired answers or single mismatches is likely due to the fact that there are many more possible multiple mismatches that productively hybridized with the template under the ligation conditions that were used than single mismatches. Nonetheless, it should be noted that although erroneous answers represented a significant fraction of the set of sequence solutions, from a computational point of view the correct 'answer' to the 'query' would have been unknown prior to the experiment and averaging the flawed answers yields a completely correct consensus sequence solution to the search algorithm.

Table 1. Fidelity of sequence solutions as a function of time and temperature

	Error rate At 16°C for 1 min	At 16°C for 1 h	At 25°C for 1 min	At 25°C for 1 h
R12	3.1	4	2.6	3.4
R6	0.6	1.6	0.1	0.8
R3	1.8	4.7	0	10.4

The data shown in Figure 3a and the corresponding data used to derive Figures 4 and 5 were used to derive relative fidelities. Each mismatch, deletion or insertion relative to the wild-type sequence solution was counted as a single error. The number of errors was then divided by the number of sequence solutions to determine the error rate. For example, the search algorithm was evaluated with the dodecamer pool at 16°C for 1 min with an error rate of 3.1 mismatches, deletions or insertions per sequence solution.

These results are comparable with those obtained from similar experiments with dodecamer pools and chemically catalyzed ligation reactions (3.9 mistakes/sequence solution at 25°C; 9). However, we had hoped that the use of an enzyme as a coupling reagent would yield a larger proportion of correct sequence solutions to the search algorithm, since the enzyme could in principle assess the 'correctness' of pairing between an oligonucleotide substrate and the query sequence. In fact, the enzyme appeared to proofread base pairings to some extent, since the fidelity of the ligation junctions, the sites at which the catalyst would have acted, were generally higher than for the interior of the query sequence (Fig. 3c). Interestingly, while the relative fidelities of individual base pairs were similar irrespective of temperature or time, the third position of the ligated dodecamers (A) seemed to be very sensitive to reaction conditions. Ligation at 25°C for 1 min produced Watson–Crick base pairs 75% of the time, while ligation at 16°C for 1 h yielded correct pairings at less than half of this frequency. A close examination of the sequence data shows that the decrease in fidelity at lower temperatures and longer times is largely due to the stabilization of G:T base pairs (for example, the data obtained following ligation at 16°C for 1 min). These results re-emphasize the known context dependence of base pairing and suggest that the methods described in this paper might also be used to generate a wealth of data regarding the thermodynamics of base pairing in different contexts.

The trends within the fidelity data (Table 1) can be readily rationalized. There is less fidelity at 16 than at 25°C. The free energy of hybridization would have been greater at 16 than at 25°C and thus mismatches would have been better tolerated at the lower temperature. There is less fidelity after ligation for 1 h than after ligation for 1 min. The correct products would have formed and been fixed in place most quickly, while incorrect products would have accumulated with time. These trends can also be observed for individual DNA base pairs (Fig. 3c). Taken together, these results indicated that the evaluation of the search algorithm and other DNA computations can potentially be optimized by adjusting the thermodynamic and kinetic parameters of the DNA computer itself.

Optimization of the search algorithm with shorter pools

Based on our initial results, we hypothesized that further changes in temperature or time parameters could yield somewhat more accurate evaluations of the search algorithm, but that truly robust changes could best be obtained by changing the DNA components. The search algorithm was therefore performed again using shorter (hexamer and trimer) random sequence pools. Shorter pools would likely yield higher fidelity sequence solutions for three reasons. First, the melting temperatures of each oligonucleotide species would be correspondingly lower and thus individual mismatches would be expected to have a proportionately larger effect on the ability of a substrate to hybridize to the query sequence at a given temperature. Second, the number of members of the random sequence pools was much smaller (4096 hexamers or 64 trimers). While the overall probability of finding correct sequence solutions would have been the same irrespective of whether hexamer or dodecamer pools were used (i.e. 4096 × 4096 = 16 777 216), the fidelity of each individual ligation event may have been higher because fewer 'incorrect' substrate:template pairings would have been assessed each time by the ligase. Consistent with this explanation, more single base mismatches were observed with hexamer pools than with dodecamer pools. Third, the distance between ligation junctions would be smaller for shorter oligonucleotide pools. Given that the enzyme ligase appeared to be enforcing complementarity at and near ligation junctions, the shorter distance between ligation

a

R12 Min. 16°

	A	B	C
W-C	ACAGACGAACGC	TCAGACTATGCG	CTGTAGGTAAGG
A01	ACAGACGAACGC	TTGGGCTATGCG	CTGTAGGTAAGG
A02	ACAGACGAACGC	TCAGTATACGCG	CTGTAGGTAAGG
A03	ACAGACGAACGC	TTGGACTGTGCG	CTGTAGGTAAGG
A04	ACAGACGAACGC	TCATACGATGCG	CTGTAGGTAAGG
A05	ACAGACGAACGC	TCAGGCTATGCA	CTGTAGGTAAGG
A06	ACAGACGAACGC	ACGGAATATACG	CTGTAGGTAAGG
A07	ACAGACGAACGC	TTAGAATATGCG	CTGTAGGTAAGG
A08	ACAGACGAACGC	TCGGTCTCTGCG	CTGTAGGTAAGG
A09	ACAGACGAACGC	TTGGTCTATACG	CTGTAGGTAAGG
A10	ACAGACGAACGC	CCGGAAGATGCA	CTGTAGGTAAGG
A11	ACAGACGAACGC	TCGGAATGCGCG	CTGTAGGTAAGG
A12	ACAGACGAACGC	TTGGATTATGCG	CTGTAGGTAAGG
A13	ACAGACGAACGC	TTGGAATATGCG	CTGTAGGTAAGG
A14	ACAGACGAACGC	TTAGCATATGCG	CTGTAGGTAAGG
A15	ACAGACGAACGC	TTGGAATATGCG	CTGTAGGTAAGG
A16	ACAGACGAACGC	TCGGAATGCGCG	CTGTAGGTAAGG
A17	ACAGACGAACGC	TCAGATTACGCG	CTGTAGGTAAGG

R12 Hr. 16°

	A	B	C
W-C	ACAGACGAACGC	TCAGACTATGCG	CTGTAGGTAAGG
B01	ACAGACGAACGC	TCTTAGTAGGCG	CTGTAGGTAAGG
B02	ACAGACGAACGC	TGTGAGTACGCG	CTGTAGGTAAGG
B03	ACAGACGAACGC	TCATTCCAGGCG	CTGTAGGTAAGG
B04	ACAGACGAACGC	CCTGAATGTGTG	CTGTAGGTAAGG
B05	ACAGACGAACGC	ACGGAGAATGCG	CTGTAGGTAAGG
B06	ACAGACGAACGC	TTCGAATATGTG	CTGTAGGTAAGG
B07	ACAGACGAACGC	TCATGATATGCG	CTGTAGGTAAGG
B08	ACAGACGAACGC	TGTGAGAATGCG	CTGTAGGTAAGG
B09	ACAGACGAACGC	CCCGCCTTTGCG	CTGTAGGTAAGG
B10	ACAGACGAACGC	TGAGACAATGTG	CTGTAGGTAAGG
B11	ACAGACGAACGC	TTTG---ACTCG	TTGTAGGTAAGG
B12	ACAGACGAACGC	TTAGGCTAAGCG	CTGTAGGTAAGG
B13	ACAGACGAACGC	TCGGA--A-GCG	TTGTAGGTAAGG
B14	ACAGACGAACGC	TGGGCCTATGCG	CTGTAGGTAAGG
B15	ACAGACGAACGC	ATAGAGTTTGCG	CTGTAGGTAAGG
B16	ACAGACGAACGC	TCAGTTTCTGCG	CTGTAGGTAAGG
B17	ACAGACGAACGC	TTAGAT-GTGTG	CTGTAGGTAAGG
B18	ACAGACGAACGC	TTTGTCTAAGCG	CTGTAGGTAAGG
B19	ACAGACGAACGC	TCTGGGAATGCG	CTGTAGGTAAGG
B20	ACAGACGAACGC	TCGGGCTAACCG	CTGTAGGTAAGG
B21	ACAGACGAACGC	TTGTACTACGCG	CTGTAGGTAAGG
B22	ACAGACGAACGC	TTGTAGGATGCG	CTGTAGGTAAGG

R12 Min. 25°

	A	B	C
W-C	ACAGACGAACGC	TCAGACTATGCG	CTGTAGGTAAGG
G01	ACAGACGAACGC	TCCGAATACGCG	CTGTAGGTAAGG
G02	ACAGACGAACGC	TTAGAGAATGCG	CTGTAGGTAAGG
G03	ACAGACGAACGC	TCATAGTAAGCG	CTGTAGGTAAGG
G04	ACAGACGAACGC	TCAGGGTATGCA	CTGTAGGTAAGG
G05	ACAGACGAACGC	TTATAATATGCG	CTGTAGGTAAGG
G06	ACAGACGAACGC	TTAGGCTATGCG	CTGTAGGTAAGG
G07	ACAGACGAACGC	TTGGATCATGCG	CTGTAGGTAAGG
G08	ACAGACGAACGC	TTTGAAGATGCG	CTGTAGGTAAGG
G09	ACAGACGAACGC	GTAGAGTATGTG	CTGTAGGTAAGG
G10	ACAGACGAACGC	TCAGATTAAGCG	CTGTAGGTAAGG
G11	ACAGACGAACGC	TCAGACTATGCG	CTGTAGGTAAGG
G12	ACAGACGAACGC	TCTGACTATGCG	CTGTAGGTAAGG
G13	ACAGACGAACGC	TCAGGCAATGCG	CTGTAGGTAAGG
G14	ACAGACGAACGC	TCAGACTAGGCG	CTGTAGGTAAGG
G15	ACAGACGAACGC	TTAGTCGATGCG	CTGTAGGTAAGG
G16	ACAGACGAACGC	TTAAACTTTGCG	CTGTAGGTAAGG
G17	ACAGACGAACGC	TTAG-TTATGCG	CTGTAGGTAAGG
G18	ACAGACGAACGC	ACAGAGTATGTG	---TAGGTAAGG
G19	ACAGACGAACGC	TC-GCCTATGC$_{T}$G	CTGTAGGTAAGG
G20	ACAGACGAACGC	TTAGAATAAGCG	CTGTAGGTAAGG

R12 Hr. 25°

	A	B	C
W-C	ACAGACGAACGC	TCAGACTATGCG	CTGTAGGTAAGG
H01	ACAGACGAACGC	TTAGCAAATGCG	CTGTAGGTAAGG
H02	ACAGACGAACGC	TACGACTATACG	CTGTAGGTAAGG
H03	ACAGACGAACGC	TCAGG-TGTGGG	CTGTAGGTAAGG
H04	ACAGACGAACGC	TCTGACTAACCG	CTGTAGGTAAGG
H05	ACAGACGAACGC	TCAGACAATGCG	CTGTAGGTAAGG
H06	ACAGACGAACGC	TCCGAATAAGCG	CTGTAGGTAAGG
H07	ACAGACGAACGC	CTAGATGATGCG	CTGTAGGTAAGG
H08	ACAGACGAACGC	TCAGAGGATGCG	CTGTAGGTAAGG
H09	ACAGACGAACGC	TTGGGCGATGCG	CTGTAGGTAAGG
H10	ACAGACGAACGC	TTGGA-TTTGCG	CTGTAGGTAAGG
H11	ACAGACGAACGC	TCTCACTATGGG	CTGTAGGTAAGG
H12	ACAGACGAACGC	TTGGATTAAGTG	CTGTAGGTAAGG
H13	ACAGACGAACGC	-CAGAGTATGCT	CTGTAGGTAAGG
H14	ACAGACGAACGC	TAAGATTATGCG	CTGTAGGTAAGG
H15	ACAGACGAACGC	TTAGTTGATGCG	CTGTAGGTAAGG
H16	ACAGACGAACGC	ATAGATAATGCG	CTGTAGGTAAGG
H17	ACAGACGAACGC	TCCGATAATGCG	CTGTAGGTAAGG
H18	ACAGACGAACGC	TAAGGGTATGCG	CTGTAGGTAAGG
H19	ACAGACGAACGC	TTGGATAGGGCG	CTGTAGGTAAGG

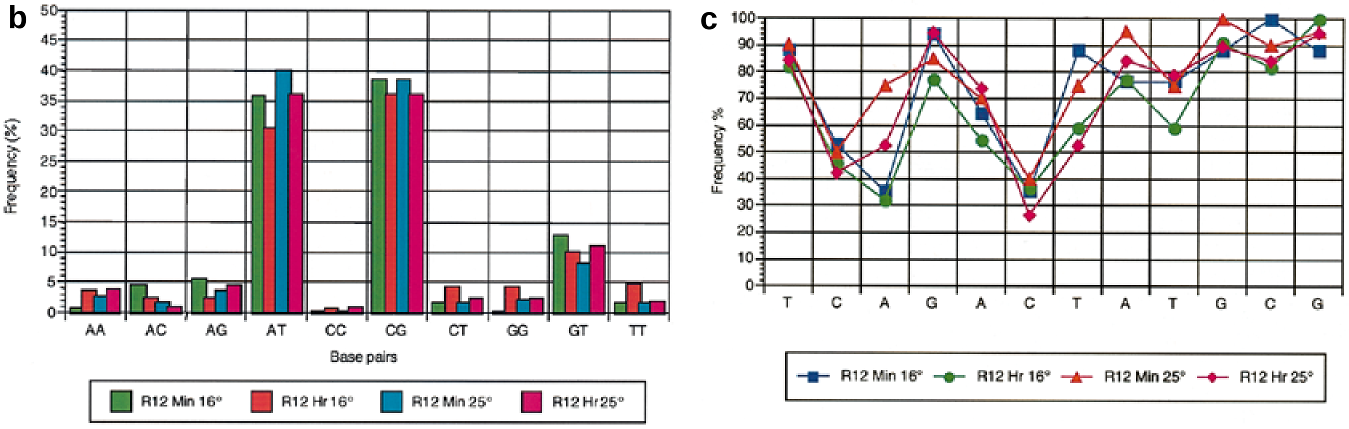

Figure 3. Fidelity of dodecamer sequence solutions. (**a**) Sequence solutions to the search algorithm described in Figure 1b. The search algorithm was carried out at either 16 or 25°C for either 1 min or 1 h. The Watson–Crick complement is shown at the top of each data set. Flanking sequences are shown in black, correctly paired nucleotides in red and mismatches in blue. Deletion variants, which likely arose from incomplete syntheses of the random pools, are indicated as dashes. (**b**) The relative frequency of predicted base pairings in sequence solutions. The data from Figure 3a is plotted in terms of predicted base pairings. Because the template region contained an equal number of each of the four bases, perfect fidelity would have resulted in 50% of the pairs being A:T and 50% being G:C. (**c**) The fidelity of pairing as a function of position. The data from Figure 3a is plotted as a function of position along the dodecamer template. The residues listed along the bottom of the graph are the correct pairing partners for the template, 5′→3′. The lines chart the frequencies of these residues under the different reaction conditions employed.

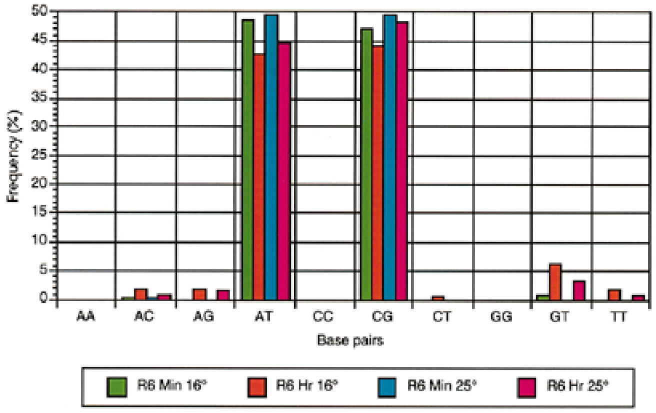

Figure 4. Fidelity of hexamer sequence solutions. Data from the evaluation of the search algorithm with hexamer pools is presented as in Figure 3b.

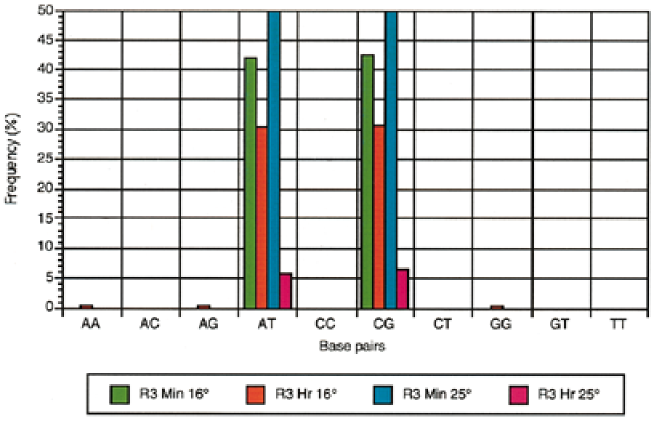

Figure 5. Fidelity of trimer sequence solutions. Data from the evaluation of the search algorithm with trimer pools is presented as in Figure 3b. Because of the large number of deletion variants that were present amongst the sequence solutions the values represented in the bars will not sum to 100%. This was also true for Figures 3b and 4, but was not readily apparent because deletion variants were a much smaller proportion of the acquired sequence solutions.

junctions might give the enzyme a greater chance to proofread each hybridization event before locking it in place.

The hypothesis that the extraction of correct sequence solutions from the potential solution set would be greatly enhanced by using shorter oligonucleotide pools as substrates was borne out in part by the experimental data. From 18 to 24 clones from each experiment were sequenced and the relative fidelities of sequence solutions selected from the hexamer and trimer pools were determined, as shown in Figures 4 and 5, respectively. A numerical summary of the fidelity data is again provided in Table 1. The observed fidelity of the sequence solutions drawn from the hexamer pool has improved greatly from those drawn from the dodecamer pool; there are now on average only 0.78 mistakes/sequence solution. The trends within the hexamer data were similar to those seen for the dodecamer pool: higher temperatures and shorter ligation times promoted fidelity. In contrast, the overall fidelity of the sequence solutions drawn from the trimer pool (4.2 mistakes/ sequence solution) was now the same or worse than those drawn from the dodecamer pool. However, this average masks the fact that when the search algorithm was carried out at 25°C for 1 min with the trimer pool it was evaluated perfectly, at least for the 18 independent clones that were examined. This result is especially impressive given that 90 individual ligation events must have occurred without mistake to generate the ensemble of clones. As before, the error rate with the trimer pool was higher at lower temperatures and longer times, but the increase in the relative number of errors was much larger than had previously been observed with the dodecamer or hexamer pools.

Artifacts of the search algorithm

The fidelity of the sequence solutions extracted from the trimer pool was decreased by the formation of artifactual side products. These artifactual side products were always present in the ligation reactions, but predominated during the evaluation of the search algorithm with the trimer pool because sequence solutions were inefficiently generated when trimers were used as substrates. Gel analysis of ligation reactions with the gapped duplex and four perfectly complementary trimers revealed much less efficient gap filling than corresponding ligation reactions with two hexamers or with one dodecamer (data not shown). In addition, many of the clones that were isolated from reactions with the trimer pool contained from none to three, rather than four, trimers spanning the dodecamer gap.

Two different unprogrammed artifacts were observed. The first artifact involved the direct ligation of the A and C fragments to one another, bypassing the programmed requirement for the insertion of a dodecamer sequence solution to span the single-stranded gap. Interestingly, all of the direct A–C ligation products lacked a cytidine residue at the 5′-end of the C oligonucleotide or, possibly, the 3′-end of the A oligonucleotide. We had previously observed the accumulation of this artifact during chemically catalyzed ligation reactions and attributed it to the formation of an alternative conformation of the gapped duplex that promoted direct A–C ligation (Fig. 6a). However, this alternative conformation is predicted to be relatively unstable (9) and direct A–C ligation products were primarily observed when enzymatic ligation was inefficient. While no direct A–C ligation products were observed with the hexamer or dodecamer pools, this artifact ranged from 11% of all ligation products with the trimer pool at 16°C for 1 min to 71% of all ligation products with the trimer pool at 25°C for 1 h. In other words, the direct A–C ligation event was a competitor of the gap filling reaction and can be viewed as an internal control for comparing the relative efficiencies of enzyme-catalyzed gap filling reactions.

The origins of the second artifact were decidedly more complex. In addition to the direct A–C ligation products described above, four clones that contained a dodecamer insert were isolated from ligation reactions with the trimer pool at 25°C for 1 min. Even though the clones were identical to one another, the dodecamer insert was not complementary to the programmed template. In fact, the alternative sequence solution that was found, 5′-ACAGACGAACGC-3′, was quite different from the expected sequence solution, 5′-TCAGACTATGCG-3′. Since it was unlikely that these four sequence solutions could have arisen spontaneously by an untemplated ligation process, it seemed probable that an alternative template may have given rise to them, just as an alternative conformation of the gapped duplex gave rise to direct

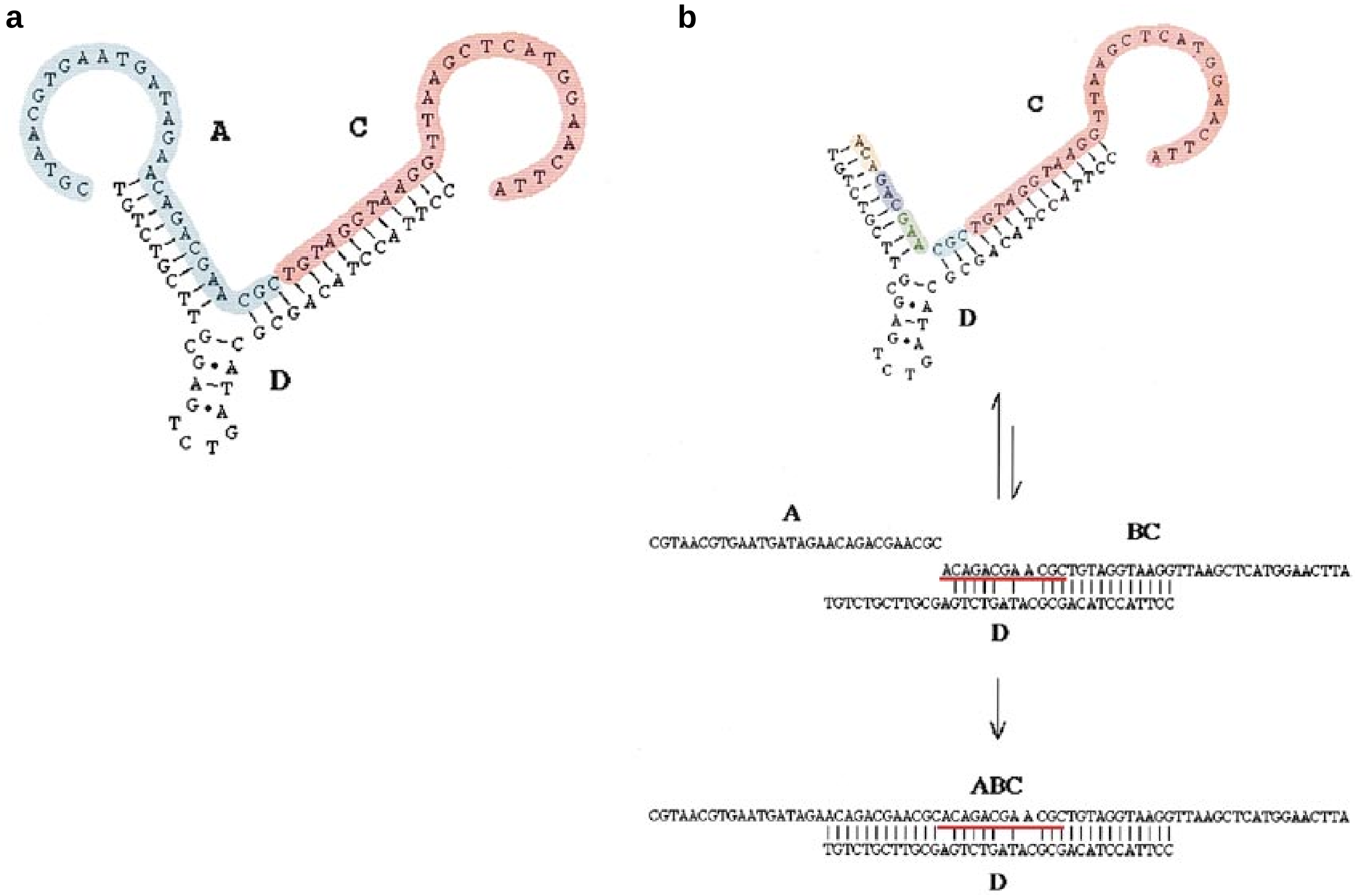
a
A
C
D
b
C
D
A
CGTAACGTGAATGATAGAACAGACGAACGC
BC
ACAGACGA A CGCTGTAGGTAAGGTTAAGCTCATGGAACTTA
TGTCTGCTTGCGAGTCTGATACGCGACATCCATTCC
D
ABC
CGTAACGTGAATGATAGAACAGACGAACGCACAGACGA A CGCTGTAGGTAAGGTTAAGCTCATGGAACTTA
TGTCTGCTTGCGAGTCTGATACGCGACATCCATTCC
D

Figure 6. Artifactual ligation products. (**a**) Proposed mechanism for direct A–C ligation. The template, oligonucleotide D, can form an alternative conformation that involves a short stem. The stem is stabilized by A:G base pairings, which are in turn stabilized at 0 relative to 25 °C (9,18). When the cytidine at the 5′-terminus of oligonucleotide C (shaded in red) is missing (as shown in this figure), the 3′-terminus of oligonucleotide A can form three adjacent base pairings. Ligation links the 5′-phosphate of oligonucleotide C with the 3′-hydroxyl of oligonucleotide A, generating a direct A–C ligation product with a missing cytidine residue. Alternatively, the cytidine at the 3′-terminus of oligonucleotide A may be missing and a direct A–C ligation product with a full-length oligonucleotide C could still be formed. (**b**) Proposed mechanism for the generation of alternative dodecamer sequence solutions. The alternative conformation of oligonucleotide D forms, as in (a). Oligonucleotide C (shown in red) binds adjacent to a correctly paired trimer (shown in blue) and ligase fixes the trimer in place. Additional trimers (shown in yellow, purple and green) bind to the region of oligonucleotide D that is normally occupied by oligonucleotide A and are again fixed in place by ligase. The 'BCD' complex then undergoes a second conformational change that enables oligonucleotide A to now hybridize adjacent to oligonucleotide D. The four ligated trimers are underlined in red in the new BCD conformation. Following ligation, an ABC ligation product with several mismatches is generated. The four ligated trimers are again underlined in red in the ABC ligation product. Again, it is also possible that the cytidine at the 3′-terminus of oligonucleotide A rather than the 5′-terminus of oligonucleotide C may be missing and an ABC ligation product could still be formed by the ligation of four trimers.

A–C ligation products. In this respect, it was suspicious that the incorrect sequence solutions once again appeared to lack a cytidine at either the 5′-end of the C oligonucleotide or the 3′-end of the A oligonucleotide. Moreover, while the dodecamer insert was not complementary to the programmed template the insert was a direct repeat of the 3′-end of oligonucleotide A. We therefore hypothesized that the same alternative conformation of the gapped duplex that gave rise to direct A–C ligation products also acted as a template for the enzyme-catalyzed ligation of four trimers, as depicted in Figure 6b. Following trimer ligation, a second conformational change allowed the A oligonucleotide to bind adjacent to the newly synthesized dodecamer and led to the formation of a full-length but incorrect sequence solution that could be amplified by PCR. The most surprising feature of both of these models is that ligase can apparently both ignore a stem–loop in the template strand (Fig. 6a) and actually directly ligate substrates presented across a gap created by this stem–loop (Fig. 6b). The direct ligation of template-bound DNA oligonucleotides presented across a single-stranded sequence gap has previously been observed using T4 RNA ligase as a coupling reagent (10).

Implications for DNA computing

The results of the evaluation of our search algorithm have both favorable and unfavorable implications for DNA computations. In the most favorable light, the number of mistakes yielded by a DNA computation would be the number of wrong answers generated per operation performed, or in this case the number of wrong oligonucleotides ligated in place per oligonucleotide considered. When the search algorithm is evaluated with a dodecamer pool there are of the order of three mistakes per sequence solution (Table 1). Since there are 5940 three mutant variants ($\{12! \div [9!\ 3!]\} \times 3^3$) of the correct sequence solution the error rate for the search algorithm carried out with a dodecamer pool could be calculated as 5940 three mutant variants/16 777 216 oligonucleotides considered or ~1 wrong oligonucleotide/2824 oligonucleotides considered. This number compares favorably with some estimates of the frequency of mistakes generated by Intels formerly flawed Pentium chip (up to 1 mistake/10^3 divisions), although not with Intels own estimates of the frequency of mistakes (1 mistake/9 × 10^9 divisions) (11,12). Alternatively, the number of mistakes yielded by a DNA computation can be estimated as the number of wrong answers produced per correct answer, rather than per operation performed. In this case, the error rate for DNA computation with a dodecamer pool would be of the order of 5940 wrong oligonucleotides ligated in place for every correct sequence solution. In practice, though, we have identified conditions for the evaluation of the search algorithm where few or no (within the limits of our sequencing data) incorrect answers were returned.

Having evaluated the search algorithm under a variety of conditions we can make cogent suggestions for the design of future DNA computations. For computations that involve hybridization, ligation and amplification, such as the HPP, Maximal Clique and addition problems so far attempted, higher ligation temperatures and shorter ligation times are likely to produce more accurate answers. Similarly, Frutos *et al*. (7) have shown that for a set of DNA 'words' designed for DNA computations that mispairings can be reduced by carrying out hybridization at 37 rather than 22°C. Alternatively, the length of the substrates for ligation can be reduced. When hexamer and trimer pools were used to evaluate the search algorithm, the fidelity of the sequence solutions was greatly increased. Finally, a proofreading or error correction function would help to reduce the divergence between sequence solutions and produce more accurate answers. The simplest error correction function would be to average the extant sequence solutions, an operation that should return the correct sequence solution as a consensus. The amplified, double-stranded DNA products could also be 'checked' against the right answer by a process of denaturation and re-hybrdiziation to an immobilized query sequence. It may even be possible to iteratively select sequence solutions and accumulate gains in fidelity, as suggested by Cukras *et al*. (5).

The use of shorter oligonucleotide substrates will not only improve the fidelity of the evaluation of algorithms but should also assist in resolving an even more fundamental problem: how to encode computationally complex problems. In general, any problem that would give pause to a supercomputer can only be resolved as a DNA computation using a large number of specifically encoded oligonucleotides. It may be possible to encode extremely diverse oligonucleotide sets either by brute force serial synthesis or by more clever synthetic strategies involving mix-and-split approaches (5) or parallel synthesis on chips (7). However, the limitations of synthesis technology will likely restrict encoded sets for DNA computations to regions that span from four to 20 residues, clustered or dispersed. For example, the overlap between oligonucleotide cities and paths in the HPP (1) was only 10 residues, the DNA word sets envisioned by Frutos *et al*. (7) would encode information in eight sequence positions and the encoded 'bits' of information in the Maximal Clique (4) and satisfiability (SAT; 5) computations were 10 and 15 residues in length, respectively. Since there are many fewer possible oligonucleotides that differ by only three to six residues from one another than oligonucleotides that differ by up to 12 or more residues from one another, the use of shorter encoded oligonucleotide sets would increase the synthetic tractability of DNA computations.

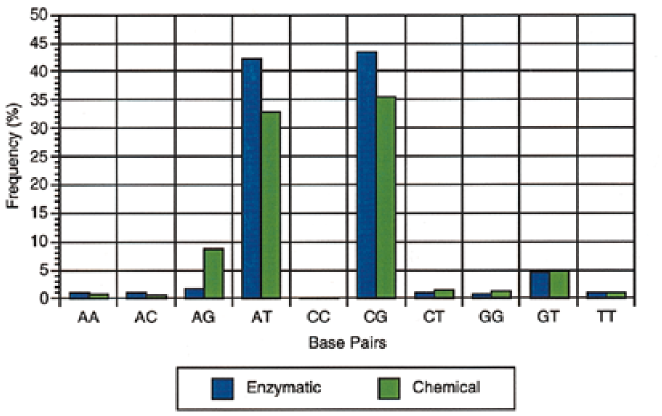

Figure 7. Comparison of the fidelity of enzymatic and chemical ligation. The data underlying Figures 3b, 4 and 5 were combined and plotted as the relative frequencies of different base pairings. Similar data for chemical ligation that have previously been reported (9) were also combined. As in Figure 5, the values represented by the bars do not always sum to 100% because of the presence of deletion variants.

If encoded oligonucleotide sets are of the order of 20 nt in length, then the molecular biology operations of hybridization and ligation will likely be carried out at the moderate temperatures favored by enzymes from mesophiles rather than the high temperatures favored by enzymes from thermophiles. Our experiments reveal that there may be inherent limitations on the fidelity of at least some mesophilic calculations. First, the fidelity of available mesophilic DNA ligases will in turn determine the fidelity of calculations. The *E.coli* ligase utilized in these studies seemed to be able to avoid or proofread at least some errors. A comparison between the results presented here and chemical ligation experiments with the same template (9) reveals that many fewer A:G base pairs were accepted by enzymatic ligation (Fig. 7). This is especially true when fidelity is examined as a function of residue position: while A:G base pairs were frequently found at ligation junctions produced by chemical ligation, Watson–Crick base pairs predominated at ligation junctions produced by enzymatic ligation, as has previously been observed for thermophilic ligases (13). However, the fidelity of the *E.coli* ligase decreased for internal base pairings. These results are consistent with reports that T4 DNA ligase can join mismatched DNA substrates, albeit inefficiently (14,15). To improve the fidelity of DNA computations, it may be possible to engineer or evolve ligase variants with improved fidelities (16). Second, even though the short encoded oligonucleotides that may dictate mesophilic DNA computation will be less likely to form secondary structures that interfere with the evaluation of algorithms than long encoded oligonucleotides, artifacts can nonetheless arise. The direct A–C ligation products and templated but incorrect sequence solutions that arose during evaluation of the search algorithm with the trimer pool are examples. Similarly, mismatched sequences that accumulated during a selection of substrates preferred by T4 DNA ligase were likely the result of an artifactual selection for sequences that could amplify well in PCR, rather than for sequences that were in fact preferred by the ligase (9,17).

Overall, this paper represents the first effort to determine whether computationally complex problems that could not be readily solved on an electronic computer might be quickly and accurately evaluated using a DNA computer. The large number of errors that accumulated during the evaluation of the search algorithm does not bode well for the evaluation of large, computationally complex problems via the molecular biology operations of hybridization, ligation and amplification. While these molecular biology operations can be optimized, the very need for optimization may imply that each new DNA computation will require not only a significant encoding effort but also a lengthy procedural optimization as well. The best example of the need for precise and likely lengthy procedural optimization was observed during the evaluation of the search algorithm with trimer pools. When the algorithm was carried out at 25°C for 1 min only correct sequence solutions were returned. In contrast, when either the time or temperature were changed almost no correct sequence solutions were returned. These conclusions do not necessarily imply that DNA computations are inherently infeasible, but may merely point towards the need for different molecular biology operations for evaluation. While evaluation of the HPP and addition problems relied exclusively on hybridization, ligation and amplification to generate correct sequence solutions and thus might be termed ‘additive’ approaches, evaluation of the Maximal Clique and SAT problems relied on either chemical or enzymatic synthesis of all possible sequence solutions followed by hybridization, cleavage and amplification to generate correct sequence solutions and thus might be termed ‘subtractive’ approaches. To the extent that the molecular biology operations of hybridization and cleavage may be inherently more faithful than the molecular biology operations of hybridization and ligation then DNA computations based on these operations may also be performed with greater fidelity.

REFERENCES

1 Adleman,L.M. (1994) *Science*, **266**, 1021–1024.
2 Guarnieri,F., Fliss,M. and Bancroft,C. (1996) *Science*, **273**, 220–223.
3 Lipton,R.J. (1995) *Science*, **268**, 542–545.
4 Ouyang,Q., Kaplan,P.D., Liu,S. and Libchaber,A. (1997) *Science*, **278**, 446–449.
5 Cukras,A.R., Faulhammer,D., Lipton,R.J. and Landweber,L.F. (1998) In Kari,L. (ed.), Chess Games: *A Model for RNA Based Computation*. Proceedings of the 4th Annual Meeting on DNA Based Computers, in press.
6 Rozen,D.E., McGrew,S. and Ellington,A.D. (1996) *Curr. Biol*., **6**, 254–257.
7 Frutos,A.G., Liu,Q., Thiel,A.J., Sanner,A.M.W., Condon,A.E., Smith,L.M. and Corn,R.M. (1997) *Nucleic Acids Res*., **25**, 4748–4757.
8 Maniatis,T., Fritsch,E.F. and Sambrook,J. (1989) *Molecular Cloning: A Laboratory Maunal.* 2nd Edn. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
9 James,K.D. and Ellington,A.D. (1997) *Chem. Biol*., **4**, 595–605.
10 Harada,K. and Orgel,L.E. (1993) *Proc. Natl Acad. Sci. USA*, **90**, 1576–1579.
11 Pratt,V.R. (1994) http://boole.stanford.edu/pub/PENTIUM/bugs
12 Pratt,V.R. (1994) http://boole.stanford.edu/pub/PENTIUM/individual. bugs/bug21
13 Barany,F. (1991) *Proc. Natl Acad. Sci. USA*, **88**, 189–193.
14 Wiaderkiewicz,R. and Ruiz-Carrillo,A. (1987) *Nucleic Acids Res*., **15**, 7831–7848.
15 Wu,D.Y. and Wallace,R.B. (1989) *Gene*, **76**, 245–254.
16 Luo,J., Bergstrom,D.E. and Barany,F. (1996) *Nucleic Acids Res*., **24**, 3071–3078.
17 Harada,K. and Orgel,L.E. (1993) *Nucleic Acids Res*., **21**, 2287–2291.
18 Patel,D.J., Kozlowski,S.A., Ikuta,S.A. and Itakura,K. (1984) *Biochemistry*, **23**, 3207–3217.