JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 10, Number 2, 2003

Pp. 215–229

DNA Library Design for Molecular Computation

ROBERT PENCHOVSKY and JÖRG ACKERMANN

ABSTRACT

A novel approach to designing a DNA library for molecular computation is presented. The method is employed for encoding binary information in DNA molecules. It aims to achieve a practical discrimination between perfectly matched DNA oligomers and those with mismatches in a large pool of different molecules. The approach takes into account the ability of DNA strands to hybridize in complex structures like hairpins, internal loops, or bulge loops and computes the stability of the hybrids formed based on thermodynamic data. A dynamic programming algorithm is applied to calculate the partition function for the ensemble of structures, which play a role in the hybridization reaction. The applicability of the method is demonstrated by the design of a twelve-bit DNA library. The library is constructed and experimentally tested using molecular biology tools. The results show a high level of specific hybridization achieved for all library words under identical conditions. The method is also applicable for the design of primers for PCR, DNA sequences for isothermal amplification reactions, and capture probes in DNA-chip arrays. The library could be applied for integrated DNA computing of twelve-bit instances of NP-complete combinatorial problems by multi-step DNA selection in microflow reactors.

Key words: DNA library, DNA computation, code design, free energy, partition function.

INTRODUCTION

Following the experiment of Adleman (1994), it has been hypothesized that, with a large quantity of DNA, bio-molecular-based computers may offer the possibility that massive parallelism could be used for solving NP complete problems in polynomial time (Gifford, 1994; Lipton, 1995). Instances of NP complete problems, such as the maximum clique problem (Quyang *et al.*, 1997) and the SAT problem (Liu *et al.*, 2000; Braich *et al.*, 2001, 2002) have been solved using DNA/DNA hybridization. A key question in DNA computing is the fidelity of the basic operations employed and the scalability of the computation as a whole (James *et al.*, 1998; Cox *et al.*, 1999; Pevzner *et al.*, 2001). When DNA/DNA hybridization is used as a basic computational operation, the accuracy of the computation will depend on the ability to discriminate between perfectly matching hybrids (the bits of the library and their complementary oligomers) and those with mismatches. In this regard, the quality of the DNA code design is playing a critical role in the fidelity of the computation. The problem of designing sets of modular RNA and DNA sequences, which hybridize in a predefined way, is fundamental not only for molecular computing but also

Biomolecular Information Processing (BioMIP), Fraunhofer Gesellschaft, Schloss Birlinghoven, D-53754 Sankt Augustin, Germany.

for many applications in contemporary molecular biology and nanotechnology such as strand selection, PCR and isothermal amplification reactions, DNA-chip arrays, ribozyme design (Tang and Breaker, 1998), and self-assembly of DNA (Winfree *et al.*, 1998).

Word design strategies for DNA-based computation have already been investigated experimentally (Braich *et al.*, 2001, 2002; Faulhammer *et al.*, 2000; Frutos *et al.*, 1997) and theoretically (Maranthe *et al.*, 2001; Deaton *et al.*, 1998). Sequence design methods applied so far use discrimination between specific and nonspecific hybridization based on Hamming and related distances, ignoring the physical property of DNA to hybridize in more complex structures.

In this paper, we present a novel approach for DNA library design using discrimination between specific and nonspecific hybridization based on the thermodynamic stability of the DNA hybrids formed. The thermodynamic stability of certain DNA/DNA hybridization is described by the free energy of the hybridization reaction. In order to deal with the probability distribution of alternative DNA/DNA duplex structures, a partition function (McCaskill, 1990) Q is computed for the formation of all possible secondary structures for a sequence, in which the first binding partner is connected via a spacer to the second binding partner. The spacer consists of artificial nucleotides (N), which are defined to have no binding properties so that no binding is possible to the spacer region (see Fig. 1). To calculate the partition function Q and the free energy $E_b = -kT \ln(Q)$ for the ensemble of structures, we use the Vienna RNA folding package (Hofacker *et al.*, 1994) (available via *www.tbi.univie.ac.at/~ivo/RNA*) and appropriate DNA parameters (Allawi and Santa Lucia, 1997; Santa Lucia, 1998). The free energy E_b corresponds to the effective equilibrium constant for an ensemble of (probability weighted) duplex structures and has to be distinguished from the minimal free energy, which describes the hybridization for one particular structure (i.e., for the minimal free energy structure). The term "free energy" denotes consequently the effective free energy E_b derived from the partition function through this work.

To optimize the set of DNA words, a random search algorithm is applied (see Materials and Methods). The algorithm is general in the sense that any word length and any set size can be chosen; additional requirements for a special experimental setup can be implemented easily. Extensive numerical

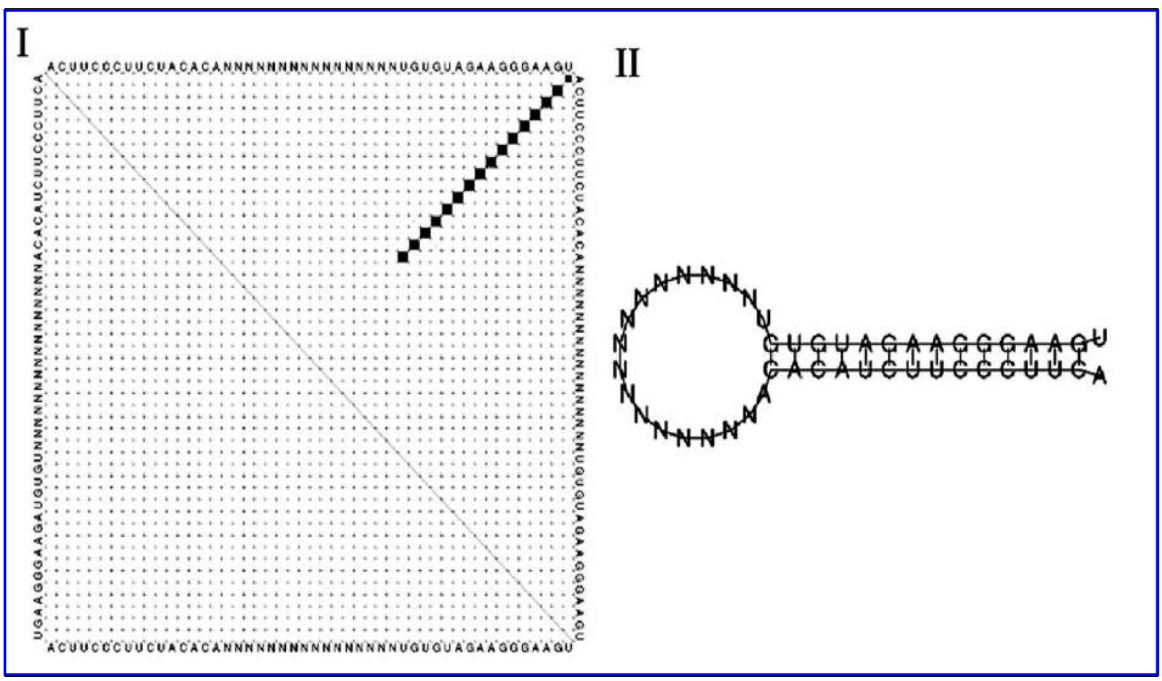

FIG. 1. Partition function. The Vienna RNA folding package is used to calculate the partition function. (**I**) The dot matrix represents base pairing probabilities. The example shows the hybridization of a word to its complement. (**II**) The dominant structure is plotted. The hybridization thermodynamics are described by an ensemble of secondary structures for a sequence, which consists of the two binding partners connected via a spacer of sixteen artificial nucleotides "N" (see the text).

optimization runs have been performed for various set sizes and for various experimental requirements. We demonstrate the approach by designing a twelve-bit library, which is assembled by DNA ligation reactions and tested experimentally by PCR, and by DNA hybridization analyses on beads and in a solution.

A bead-based approach for multistep DNA selection in steady-flow microreactor modules has been proposed previously (McCaskill, 2001) as well as an immobilization of DNA to beads (Penchovsky *et al.*, 2000). The chemistry for such multistep DNA selection under isothermal conditions by changing the pH of the solutions has been already demonstrated by cascadable hybridization transfer of specific DNA in microreactor selection modules (Penchovsky and McCaskill, 2002). Performing hybridization and denaturation steps at one temperature gives an opportunity for integration of more selection modules on one wafer than using a temperature gradient. The kinetics of DNA hybridization on beads in flow conditions (Penchovsky and McCaskill, 2002; Fan *et al.*, 1999) is much faster than that at stationary conditions (Stevens *et al.*, 1999), which gives additional advantages in using microflow reactors for multistep DNA selection.

MATERIALS AND METHODS

DNA library constraints

The conditions to be fulfilled by the DNA library presented and tested in this work are given as follows:

1. It is a twelve-bit DNA library built from 24 different words (bits) representing a "one" or a "zero" at twelve different positions.
2. The library words contain A's, T's, C's, but no G's (Mir, 1999). Avoiding the presence of G's in the sequences restricts their variability but simultaneously reduces significantly the stability of possible secondary structures and the hybridization among the library sequences (Braich *et al.*, 2001, 2002; Faulhammer *et al.*, 2000).
3. The words are 16 nt deoxyoligonucleotides. This length is a reasonable choice for primer binding sites and guarantees a large pool of $3^{16} \approx 4.3 \times 10^7$ different sequences.
4. The occurrences of four or more consecutive identical nucleotides in the words are avoided. The presence of long homopolymer tracts in the library sequences could be a reason for the (kinetically favored) formation of secondary structures. A similar constraint (not more than five consecutive identical nucleotides) is applied by Braich *et al.* (2001, 2002).
5. Runs of three consecutive C's at either the $5'$ or the $3'$ ends are avoided because that could influence the hybridization kinetics.
6. The free energy gap between the weakest specific hybridization and the strongest nonspecific hybridization within the word set should be as large as possible in order to reduce to minimum any nonspecific hybridization reaction between the DNA library sequences and the capture probes.
7. The melting points (T_m) of the words have to be in a restricted range of $\pm 1.5°C$ because we want to achieve uniform hybridization yield for all capture probes under identical conditions. To calculate the T_m values, the nearest neighbor approximation (Wetmure, 1991) and the thermodynamic data of Breslauer *et al.* (1986) and Allawi *et al.* (1997) were used.

Algorithm used for DNA library design

The design of the library was performed in two steps (A and B). In step A, the binding properties of a set of words $\{w_i, i = 1, 2, \ldots, N\}$ were optimized by a random search algorithm. In step B, the set of words obtained was utilized to construct a twelve-bit DNA library. Here we take into account the possible hybridization reactions of capture probes to sequence regions produced by the ligation of adjacent words in the library. For this purpose, the words in the set were ordered according to the value (either one or zero) and position ($i = 1, 2, \ldots, N/2$) in a bit string. For words dedicated to a bit, we employ the notation $\{V_i^n; i = 1, 2, \ldots, N/2; n = 0, 1\}$ where the upper index indicates the value of the bit and the lower index gives the position (i.e., the presence of the oligomer V_7^1 in a concatenation of words means that the bit number 7 in the bit string has the value 1).

The random search algorithm for step A can be described as follows:

(A1) Generate a set of random words w_i, $i = 1, 2, \ldots, N$ over the alphabet {A, T, C}.

(A2) Compute the N free energies E_b for the hybridization reactions of all words w_i to their Watson–Crick complements. For each word, compute the spectra Σ_I of (effective) free energies for all possible binding reactions to other words in the set and to their complements. Compute the melting temperatures T_m for all words in the set.

(A3) Choose a word w_j randomly from the set.

(A4) Generate a new random word w_r over the alphabet {A, T, C}.

(A5) Check the sequence w_r for the occurrence of four or more consecutive identical nucleotides as well as for runs of three consecutive C's at either the 5′ or the 3′ ends. Reject the sequence w_r if one of these subsequences was observed and go back to step (A4).

(A6) Compute the melting temperature T_m for the word w_r. If the melting temperature is outside of the range of melting temperatures of the previous set, go back to step (A4).

(A7) Compute the free energy E_b for the hybridization reaction of word w_r to its complement. If this binding is thermodynamically weaker than the weakest specific binding for the previous set, go back to step (A4).

(A8) Compute all free energies for the hybridization reaction of the word w_r to the word set $\{w_i, i = 1, 2, \ldots, N\}/\{w_j\}$ and the corresponding complements. If one of these nonspecific bindings is stronger than the strongest nonspecific binding for the previous set, go back to step (A4).

(A9) Replace the word w_j by the word w_r and go back to step (A2).

The algorithm is halted when the gap between the free energies E_b for the specific binding and the free energies Σ_I for the nonspecific binding has converged to a constant. The words in the library were ordered by the following algorithm (step B):

(B1) Generate all possible concatenations of two words $\{w_i : w_j, i, j = 1, 2, \ldots, N, i \neq j\}$ for the set of words optimized in step A.

(B2) For each such concatenation $w_i : w_j$, calculate the free energies for all (nonspecific) hybridization reactions to words from the set $\{w_k, k = 1, 2, \ldots, N, k \neq i, k \neq j\}$ and their complements. Save for each concatenation the free energy I_b for the strongest nonspecific binding.

(B3) Determine the concatenation $w_i : w_j$ with the weakest such nonspecific binding I_b and set $V_0^1 = w_i$ and $V_1^1 = w_j$.

(B4) Find the concatenation $w : V_1^1$ ($w \in \{w_k, k = 1, 2, \ldots, N\}/\{V_0^1, V_1^1\}$) with the weakest nonspecific binding I_b; set $V_0^0 = w$.

(B5) Search for a word $w \in \{w_k, k = 1, 2, \ldots, N\}/\{V_0^1, V_0^0, V_1^1\}$ that gives the weakest binding for the two concatenations $V_0^1 : w$ and $V_0^0 : w$; set $V_1^0 = w$.

The search process for the next word w can proceed iteratively for the concatenations $V_i^1 : w$ and $V_i^0 : w$ ($i = 1, 2, \ldots, 23$) to determine the remaining words V_i^1, V_i^0; $i = 2, \ldots, N/2$.

Having determined the word order in the library, we permuted randomly chosen word pairs in order to reoptimize the overall (average) binding properties and to maximize the slide mismatches between each possible concatenation ($V_i^n : V_{i+1}^m$; $i = 1, 2, \ldots, N/2$; $n, m = 0, 1$) and all other words in the set ($w \in \{w_k; k = 1, 2, \ldots, N\}/\{V_i^m, V_{i+1}^n\}$) and their complements.

Finally, four sequences (111111111111, 000000000000, 101010101010, 010101010101) of the library, representing all ligation subsequences, were tested for slide mismatches with all words. In addition, no word should have a run of more than seven consecutive nucleotides identical to any combination of the other 23 words in the library.

Oligodeoxynucleotides and chemicals used

The oligomers were obtained from IBA-NAPS (Göttingen, Germany). All of them were purified to HPLC grand. The capture probes (see Table 1) were 5′ amino-labeled using a C6 linker and had an additional

TABLE 1. ALL WORD SEQUENCES (BITS) WITH THEIR COMPLEMENTARY STANDS (CAPTURE PROBES) ARE SHOWN IN THE TABLE[a]

	Bits	*5′–3′ Word sequences*	*5′–3′ Capture probe sequences*	*Melting points*	*Melting points*
1	1.1	CCATCACTACCTTCAT	ATGAAGGTAGTGATGG	45.3	46.8
2	1.0	TCCTCTATCATCCTCA	TGAGGATGATAGAGGA	46.6	46.5
3	2.1	TCCCTATTCACTCTCT	AGAGAGTGAATAGGGA	44.6	46.7
4	2.0	CACACCTCAACTTCTT	AAGAAGTTGAGGTGTG	44.9	48.0
5	3.1	ACTTCCCTTCTACACA	TGTGTAGAAGGGAAGT	44.8	48.1
6	3.0	CACCATCCTTATCTCA	TGAGATAAGGATGGTG	46.7	46.4
7	4.1	TCTCTCAATCCACTTC	GAAGTGGATTGAGAGA	45.6	46.6
8	4.0	TACAATCCCACACTTT	AAAGTGTGGGATTGTA	45.6	46.6
9	5.1	TCTCTTCCTCTTACCA	TGGTAAGAGGAAGAGA	45.8	47.0
10	5.0	TCATACCTAACTCCCT	AGGGAGTTAGGTATGA	44.6	46.7
11	6.1	CTCATCTTAACCACCT	AGGTGGTTAAGATGAG	44.7	46.7
12	6.0	ACCATTACTTCAACCA	TGGTTGAAGTAATGGT	45.6	46.6
13	7.1	TTCTACAACCTACCCT	AGGGTAGGTTGTAGAA	44.4	47.3
14	7.0	TCCAACTTAACACTCC	GGAGTGTTAAGTTGGA	45.3	47.3
15	8.1	ACCTTTACCCTATCCT	AGGATAGGGTAAAGGT	46.2	47.1
16	8.0	ACACCCTAACAATCAA	TTGATTGTTAGGGTGT	45.6	46.6
17	9.1	CACCCATTCCTAATAC	GTATTAGGAATGGGTG	45.4	45.2
18	9.0	TCCTACACAAACATCA	TGATGTTTGTGTAGGA	43.8	46.3
19	10.1	ATTCTCACTCACAACC	GGTTGTGAGTGAGAAT	44.6	47.8
20	10.0	ACCACTCCAATAACTC	GAGTTATTGGAGTGGT	44.2	47.0
21	11.1	TCCTACTCTCCAATCA	TGATTGGAGAGTAGGA	46.4	47.1
22	11.0	TCTTTCACACATCCAT	ATGGATGTGTGAAAGA	46.3	46.7
23	12.1	ACACCATTTCACCTAA	TTAGGTGAAATGGTGT	45.6	46.6
24	12.0	ACACTAATCCTCCAAC	GTTGGAGGATTAGTGT	44.2	47.0

[a]Each capture probe is 5′ amino-modified by a C6 linker and has a 12 dT spacer at the 5′ end. The melting points for all words are calculated thermodynamically for 50 mM NaCl and 1 μM oligomer using two different thermodynamic data. The highest and the lowest melting points are 46.7–43.8 = 2.9°C and 48.1–45.2 = 2.9°C (without taking into account the GC and AT init. w/term.) and according to Breslauer *et al.* (1986) and Allawi and Santa Lucia (1997).

linker of 12 dT on the 5′ end because of our intention to use them for a DNA hybridization on beads. The DNA library was split into four oligomers (see Fig. 2) synthesized by a mix and split procedure (Faulhammer, 2000). All other chemicals were purchased from SIGMA (Deisenhofen, Germany) if not mentioned otherwise.

Enzymatic ligation and DNA purification

The ligation reactions were carried out at a concentration of ligated oligomers of 20 μM each. The antisense oligomers are used at a concentration of 35 μM. Tsc thermostable DNA ligase (Roche, Mannheim, Germany) was used at a concentration of 0.2 U/μl. The oligonucleotides were denatured at 93°C for 3 min and cooled to 25°C over 10 min. The ligation reactions were carried out for 2 hours at 30°C in a gradient thermoblock (Biometra, Göttingen, Germany) in the presence of 100 mM Tris-HCl pH 7.5 (at 25°C), 10 mM NaCl, 20 mM KCl, 10 mM $MgCl_2$, 0.1% Nonidet P40 (v/v), 0.5 mM NAD, and 1 mM DTT in a volume of 50 μl. The ligation products were separated from unligated oligonucleotides by 12% denaturing polyacrylamide gel electrophoresis in the presence of 63% urea and 1× TBE buffer and stained in a 1× SYBR Green Two dye (Molecular Probes, Eugene, OR). The ligated products were eluted from the gel as described by Sambroock *et al.* (1989).

PCR amplification and melting temperature estimation

The PCR experiments were performed by a gradient thermoblock from Biometra. The amplification program started with an initial denaturation at 93°C for 3 min followed by 28 cycles of denaturation at

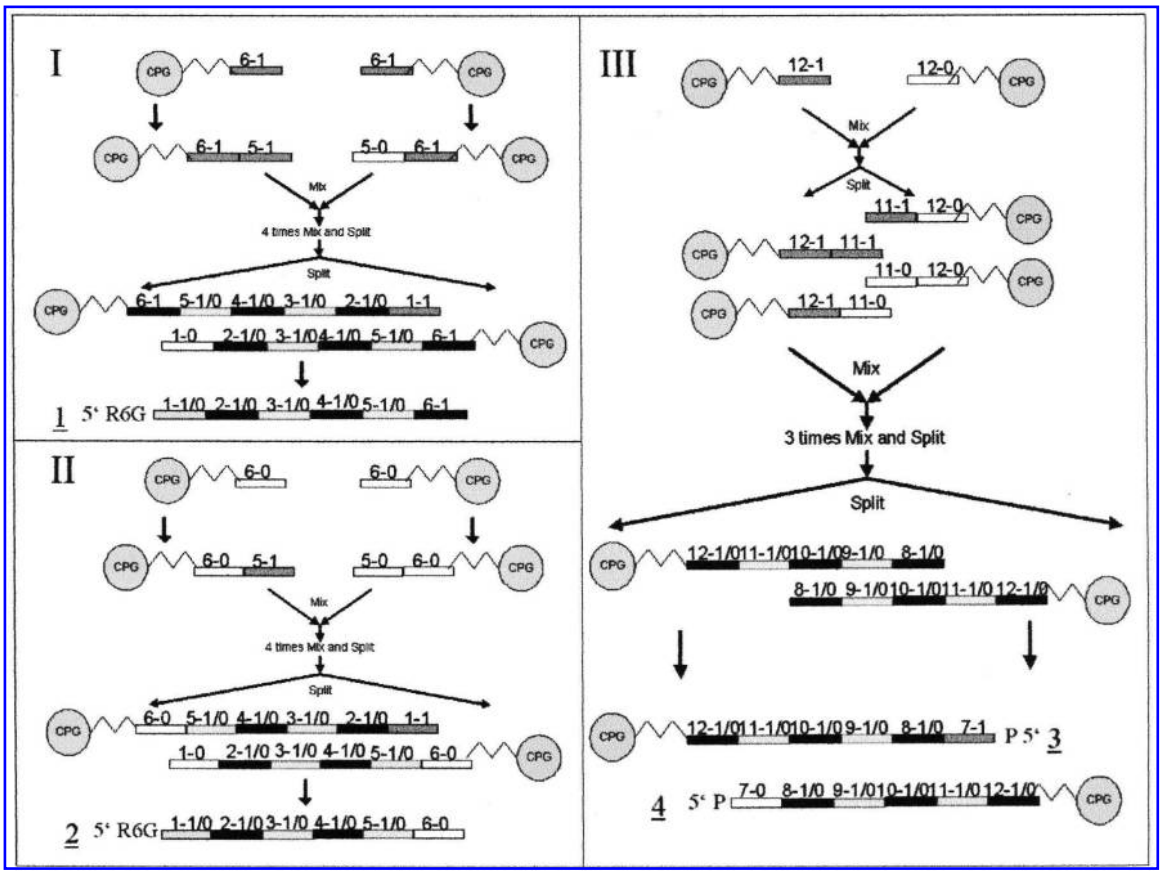

FIG. 2. Synthesis of the DNA library. The library was synthesized by four oligodeoxynucleotides each 96 nt long by a mix-and-split procedure employed by Faulhammer *et al.* (2000) Oligomer number 1 (see **I**) contains all combinations among the first five bits (1–5) and has value "one" in the sixth bit only. Oligomer number 2 (see **II**) is identical in the first five bits to number 1 but has value "zero" in the sixth bit only. Oligomer number 3 (see **III**) contains all combinations among the last five bits (8–12 bits) and has value "one" in the seventh bits, and the fourth oligomer (see **III**) is identical in the last five bits to the third one but has value "zero" in the seventh bit only.

93°C for 30 sec, annealing at 53°C for 40 sec and elongation at 72°C for 30 sec. The Taq polymerase was purchased from Qiagen (Hilden, Germany) and used in a concentration of 0.1 U/μl. The amplification reactions were carried out in a 1× incubation buffer in the presence of 1 μM sense primer, 0.5 μM antisense primer (capture probes—see Table 1), 1 pM single strand DNA template, and 250 μM dNTP's mixture. The PCR products were analyzed on 8% nondenaturing gels in the presence of 1× TBE buffer. The gels were stained in a 1× SYBRGreen One dye (Molecular Probes, Eugene, OR). The gel pictures were obtained by a gel documentation system from Biozyme (Hess, Oldendorf, Germany). All gel pictures were inverted for better readability.

In order to determine the melting temperature of the words, a real-time PCR detection system iCycler iQ from BioRad (Hercules, CA) was used. Melting curves were obtained in a 10 mM sodium phosphate buffer, pH 7.2 in the presence of 50 mM or 100 NaCl, 1× SybrGreen One dye, and 2 μM DNA strands in a final volume of 50 μl. The melting curve program started with an initial denaturation at 95°C for 3 min followed by 160 repeats, each for 1 min, with a temperature decrease of 0.5°C per repeat.

DNA immobilization and hybridization on paramagnetic beads

We immobilized 5′ amino-modified deoxyoligonucleotides at a concentration of 10 μM on 5 mg, 15 μm carboxyl coated beads (Micromod, Rostock, Germany) in the presence of 50 mM EDC, 100 mM MES

buffer pH 6.1 and 100 mM NaCl in a volume of 200 μl. The reaction was incubated for 3 h at 25°C under continuous shaking.

The DNA hybridization reactions were performed on 2 mg beads in the presence of 5 μM oligomers in 500 mM tris-borate buffer pH 8.3 (at 25°C) and 50 mM NaOH in a volume of 100 μl. The hybridization solution was made by mixing equal volumes of 1 M tris-borate buffer pH 8.3 and 100 mM NaOH solution. The pH value of the solution obtained did not change more than 0.1 units at 25°C. Those solutions previously allowed performing multistep selection by repeated DNA hybridization, denaturation, and rehybridization of denatured DNA under isothermal conditions on beads placed in microreactor selection modules (Penchovsky and McCaskill, 2002). The hybridization reactions were incubated at 36°C for 1.5 hours under continuous shaking. After each hybridization, the beads were washed twice with a hybridization buffer at 36°C over 5 min under continuous shaking. The hybridized DNA on the beads was denatured by heating at 95°C for 4 min. The concentration of denatured DNA was estimated by UV spectroscopy with a Cary 3E photometer (Varian Inc., Walnut Creek, CA) at 260 nm as described by Sambroock *et al.* (1989). When DNA was fluorescently labeled, its concentration was additionally estimated by a spectrofluorimeter FluorMax-2 (Instruments S.A., Inc., Edison, New Jersey) by using a standard titration curve.

RESULTS

Results of the algorithm

For random 16-oligomers over the alphabet {A, T, C}, the average duplex binding energy is $\langle E_b \rangle = -17.2 \pm 1.5$ kcal/mol, and the average melting temperature is $\langle T_m \rangle = 39.4 \pm 4.4$°C for 50 mM NaCl and 1 μM strand concentration in total ranges of 13 kcal/mol $\leq E_b \leq$ 23 kcal/mol and 25°C $\leq T_m \leq$ 55°C, respectively. For sets of 24 random words, the free energy of the strongest nonspecific binding has approximately a value of -11.3 kcal/mol. The gap in the free energy between specific and nonspecific hybridization in such a random set is on the order of $\delta G = 2$ kcal/mol. As the number of words increases, δG decreases and becomes zero for large set sizes (more than 100 random words). In that case, it is not possible to distinguish between specific and nonspecific DNA hybridization.

Starting with an initial set of 24 random words, step A of the algorithm has been repeated several times (more than ten) for various initial random sets (also for slightly different combinatorial constraints). Convergence of the properties of the word set has usually been archived for less than one thousand successful replacements (step A9). The free energy gap δG turned out to be rather independent of the choice of the initial word set. Several million random words have been tested (step A7/A8) in each run.

As a result of step A, the free energies for correct hybridizations converge to values within the range -20.5 kcal/mol $\leq E_b \leq -19.0$ kcal/mol, whereas the stability of mismatched DNA duplexes decreases as indicated by the free energy value of -10.1 kcal/mol for the strongest nonspecific hybridization. The free energy gap increases by a factor of more than four to a value of $\delta G = 8.9$ kcal/mol. The final melting temperatures of the words differ by 3°C. The average number of C's is 6.7 (42%) and the average number of A's and T's is 4.5 (28%), and 4.8 (30%), respectively. It is an interesting result that the word set optimized according to the thermodynamic stability of all possible binding reactions shows favorable mismatch properties (see Table 2). The contrary, however, is not true. For a DNA library of the same size, at least eight mismatches and an identical CG content of the words did show a lower thermodynamic discrimination between specific and nonspecific DNA hybridization than did the presented library. Every word of the current 12-bit DNA library has at least five mismatches with all other words or their complements (denoted by "reverse complementary slide matches"). This number may be compared with mismatch properties of the sets optimized by Frutos *et al.* (1997) to maximize the Hamming distance. For their set of 16-oligomers (set size 108), for example, the minimal number of mismatches is four.

In step B, the order of the 16-oligomers in our library was determined. We tested the four sequences (111111111111, 000000000000, 101010101010, 010101010101), covering all word boundary subsequences in the library, against all single words for slide complementary mismatches and for slide reverse mismatches.

TABLE 2. SLIDE MATCHES (S^C) AND REVERSE COMPLEMENTARY SLIDE MATCHES (S^R) FOR THE OPTIMIZED SET OF 24 WORDS[a]

Matches:nt	S^C	S^R	*Total*
16	24	0	24
15	0	0	0
14	0	0	0
13	0	0	0
12	0	0	0
11	2	0	2
10	34	0	34
9	66	0	66
8	186	1	187
7	208	0	208
6	54	40	94
5	2	184	186
4	0	285	285
3	0	66	66
2	0	0	0
1	0	0	0

[a]Binding over sixteen consecutive base pairs is possible for the hybridization for each word to its complement resulting in a value of $S^C = 24$ for 16 nt matches. Additional high number of matches indicates either possible nonspecific hybridizations (S^C) or a high probability of secondary structures in the library (S^R).

The results are listed in Table 3. Using this word order, the mismatch properties of the library is not worsened by the concatenations of two words. Every word and each capture probe has at least five mismatches with all possible concatenations of words in the library. Neither a word nor a capture probe has a run of more than seven consecutive nucleotides with a possible concatenation (to verify this, custom software was used). The strongest nonspecific binding of the library to a capture probe is associated with a free energy of −12.2 kcal/mol.

TABLE 3. SLIDE MATCHES (S^C) AND REVERSE COMPLEMENTARY SLIDE MATCHES (S^R) FOR ALL 24 WORDS TO THE LIBRARY SEQUENCES 111111111111, 000000000000, 101010101010, AND 010101010101[a]

Matches: nt	*Homology: %*	*Total* S^C	*Total* S^R	*Total* $S^C + S^R$
16	100	48	0	48
15	94	0	0	0
14	87	0	0	0
13	81	0	0	0
12	75	0	0	0
11	69	58	0	58
10	62	243	0	243
9	56	692	0	692
8	50	1498	5	1503
7	44	2517	31	2548

[a]The four sequences represent all ligation subsequences in the whole library. The four library sequences contain 48 (4 × 12) words and a capture probe hybridization over 16 bp is possible for these 48 words ($S^C = 48$ for 16 nt matches). Additional high number of matches indicates either possible nonspecific hybridizations of capture probes to one (or several) of these library sequences (S^C) or a high probability of secondary structures (S^R).

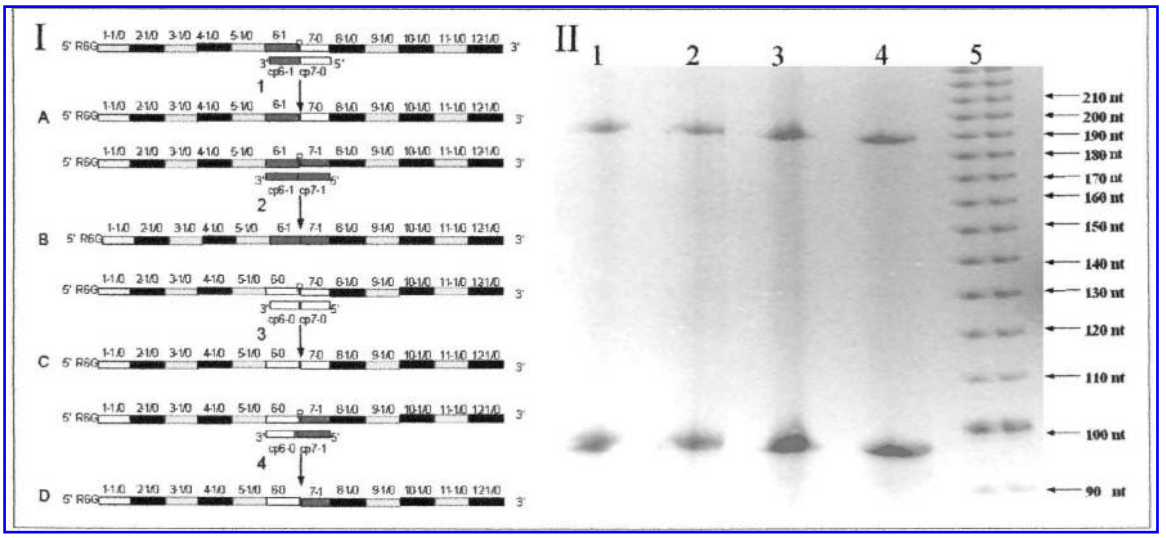

FIG. 3. Assembly of the DNA library. The library was assembled by four ligation reactions presented in scheme **I**. The products of the ligation reactions are analysed on a 12% denaturing polyacrylamide gel shown in the picture **II**—see lanes 1, 2, 3, 4. A denatured 10 bp ladder (Invitrogen GmbH, Karlsruhe, Germany) was run in the 5th lane. The ligated products are approximately 192 nt long.

Synthesis and assembly of the library

The quality of the synthesized DNA oligomers plays a critical role for the accuracy of the DNA computation. The synthesis of deoxyoligonucleotides up to 140 nt long is a standard procedure. In our case the library sequences have a length of 192 nt. DNA with that length is not very stable in a 28% ammonium hydroxide solution (mainly as a result of depurination), which is used for cleavage of the synthesized DNA oligomers from the solid support. In order to reduce this effect, we have divided the whole library into four deoxyoligonucleotides each 96 nt long (see Fig. 2). All DNA oligomers were synthesized by the mix-and-split procedure applied by Faulhammer *et al.* (2000). The first half of the library (from the first to the sixth bit) is represented by the oligomers numbers one and two (see Figure 2:I and II). They are identical in the first five bits. The oligomer with a number one has only a value "one" at the sixth bit as that with a number two has only a value "zero" at the same position. The second half of the library is represented by the oligomers three and four (see Figure 2:III). They contain all possible combinations from the eighth to the eleventh bit. The oligomer number three possesses only value "one" at the seventh as that with number four has only "zero" at the seventh position. The first two deoxyoligonucleotides are 5′ rhodamine 6G labeled, as the second two are 5′ phosphate modified.

The complete library was assembled by four ligation reactions (see Materials and Methods) schematically shown in Figure 3:I. All four ligation reactions take place between two different oligomers, one representing the first half of the library (oligomers number one or two) and another for the second half (oligomers number three or four). It was used one 32 nt oligomer, which was a complement to the bits (words) at the sixth and seventh positions (see Fig. 3), and different for each reaction. The results of the ligation reactions are analyzed on a denaturing polyacrylamide gel shown in Figure 3:II. As one can see from the gel picture, the ligated products have the expected length.

Testing integrity and accuracy of the library

In order to confirm the integrity of the assembled DNA library and to test its accuracy, twenty-two different PCR amplifications were performed (see Figs. 4 and 5) under identical conditions (see Materials and Methods). In all PCR experiments, the assembled DNA library was used as a template and the bit V_1^1 was used as a sense primer. The capture probes to the words from the second to the twelfth position were used as antisense primers in the twenty-two different PCR experiments. The amplified products with

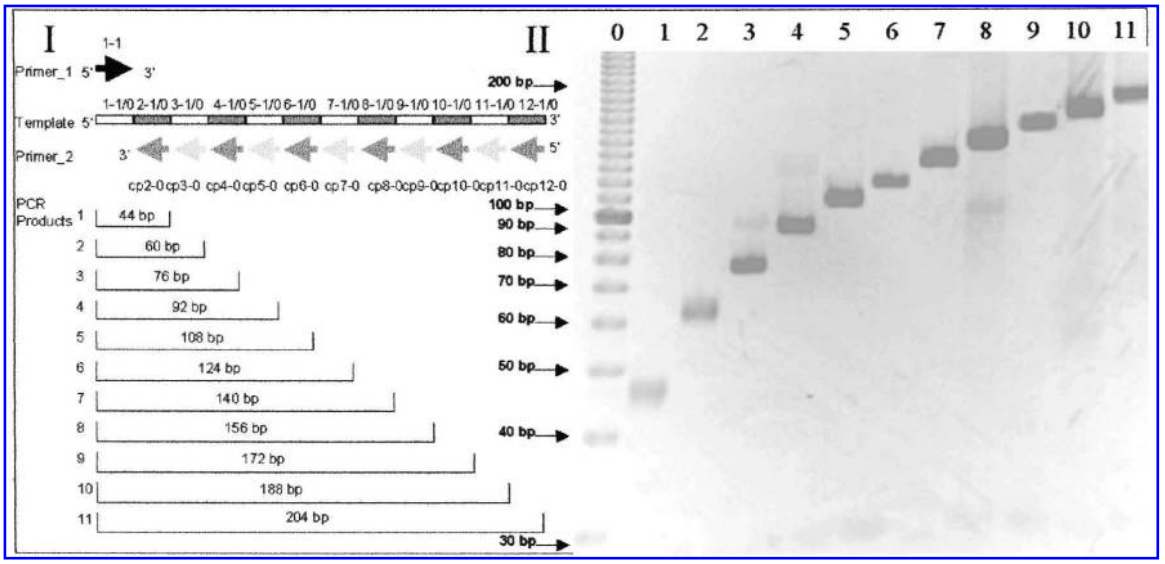

FIG. 4. Testing the integrity and the fidelity of the library by PCR amplifications with "zero" capture probes. Eleven different PCR amplifications with DNA library are presented in scheme **I**. The word V_1^1 is used as a sense primer in all reactions. As an antisense primer, the capture probes with value "zero" from the second to the twelfth position are used in eleven reactions. All reactions are performed under identical conditions (see the text). The PCR products were analyzed on an 8% nondenaturing polyacrylamide gel shown in picture **II**—see lanes from 1 to 11. On lane 0, a 10 bp nondenatured (see Fig. 3 legend) ladder is run as a marker.

"zero" capture probes were analyzed on an 8% nondenaturating polyacrylamide gel shown in Figure 4:II. As one can see from the gel picture, all 11 PCR products have the expected length (see Figure 4:I and II). No significant unexpected amplified products were observed.

Similar results were obtained for the PCR experiments with the "one" capture probes, as shown in Fig. 5. The PCR products also had the expected length with only one significant exception observed for the capture probe V_7^1 (see Fig. 5:I and II).

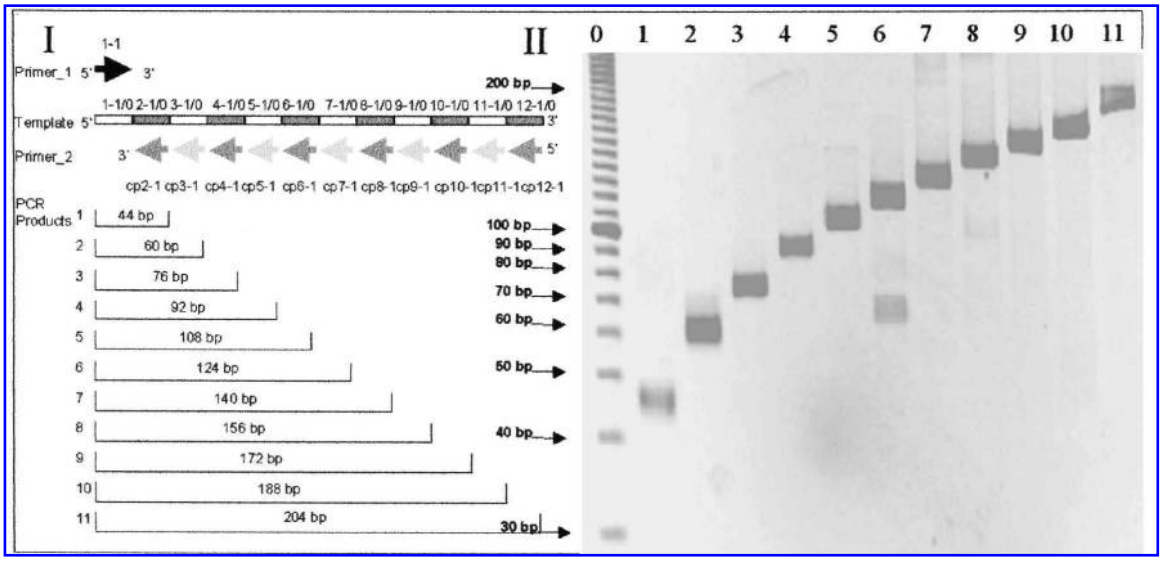

FIG. 5. Testing the integrity and the fidelity of the library by PCR with "one" capture probes. Eleven different PCR amplifications similar to those from Fig. 4 are presented. In difference to Fig. 4, the capture probes with value "one" were used as an antisense primer.

In order to localize the unexpected amplified product, four additional PCR experiments were performed. As a template, four sequences different for each reaction were employed. The capture probe V_7^1 was applied as an antisense primer in all reactions. In the first PCR experiment, the sequence 111111111111 was used as a template and the word V_1^1 as a sense primer. In the second reaction, the sequence 101010101010 was employed as template together with the same sense primer as in the first reaction. The template sequences 000000000000 and 010101010101 were used in the third and fourth reactions, respectively, and the word V_0^1 was used as a sense primer in both reactions. The PCR products were analyzed on a nondenaturing polyacrylamide gel shown in Fig. 6:I. As one can see from the gel picture, the longer amplified products with the templates 111111111111 and 101010101010 have the expected length of 124 bp. The unexpected PCR products have a length of about 70 bp in both cases. In the other two PCR experiments, no amplified products were observed. Based on these results, we conclude that the unexpected PCR products in the first and the second reactions should involve a mispriming between a block of 7 nt on the 3′ end of the capture probe V_7^1 and the template region from 40 to 46 nt which is identical in the sequences 111111111111 and 101010101010 shown in Figure 6:II. If this explanation is correct, in both cases the unexpected product should be 68 bp long as was found approximately in that gel analysis. The additional amplified product in the second reaction is more abundant than that in the first reaction due to four additional matches close to the 3′ end of the capture probe V_7^1 and the sequences 101010101010 (see Figure 6:II).

Two different DNA/DNA hybridization analyses were made with the capture probe V_7^1 immobilized on paramagnetic beads as described in Materials and Methods. We decided to focus on that capture probe because of the PCR results described above. In the first type of hybridization reactions, the second half

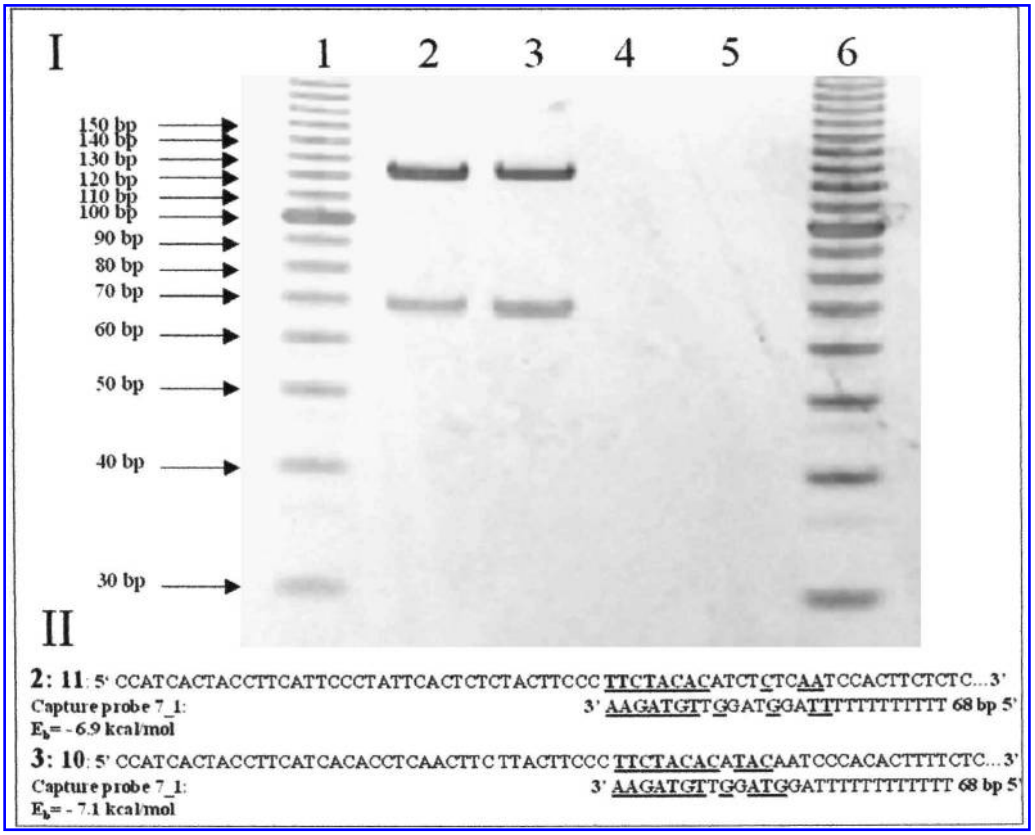

FIG. 6. Finding the origin of the wrong signal with capture probe V_7^1. **I.** Four PCR reactions are made with the capture probe V_7^1. In lane 2 the sequence 111111111111 was used as a template, in lane 3: 101010101010, lane 4: 000000000000, and in lane 5: 010101010101. In lanes 1 and 6 a 10 bp ladder was run. The wrong products in lanes 2 and 3 are a little bit shorter than 70 bp. **II.** There is only one place, which starts in both sequences from the 40th nucleotide, that has a free 3′ end and produces amplification products of 68 bp as shown.

of the library containing "one" at position seven was used. The second type of hybridization analysis was made with the first half of the library. Beads without immobilized DNA have been used as a control for nonspecific DNA attachment. The amount of denatured DNA from the beads was estimated by UV-spectroscopy and titration of the fluorescent signal when the DNA was R6G labeled.

In the first case, the hybridization yield was 41 ± 4 pmol DNA/mg beads. In the second hybridization experiment, the hybridization yield was 5 ± 2 pmol DNA/mg. The nonspecific DNA attachment to beads was 3 ± 1 pmol DNA/mg beads. Those results indicate that the nonspecifically amplified products with the capture probe V_7^1 do not provide a significant problem for the hybridization experiments with the DNA library.

In order to check the accuracy of the theoretically predicted margin between the highest and lowest melting point of the words, melting curves were obtained for each word for two different salt concentrations (50 mM and 100 mM NaCl) as described in Materials and Methods. The sequence 111111111111 was used as a template for obtaining the melting curves of the capture probes with value "one." The sequence 0000000000 was used with the "zero" capture probes. The margin for both salt concentrations was found to be ±2.5°C, which is about 1°C more than the theoretically predicted one (see Table 1) but still close enough to guarantee identical hybridization conditions for all capture probes.

Additional melting curves were obtained with the entire library in the presence of capture probes V_4^1 and V_7^1 or in absence of a capture probe (see Fig. 7). The relatively low fluorescent signal of the library-melting curve in the absence of a capture probe indicates a low level of hybridization among library sequences and a low level of secondary structures formed.

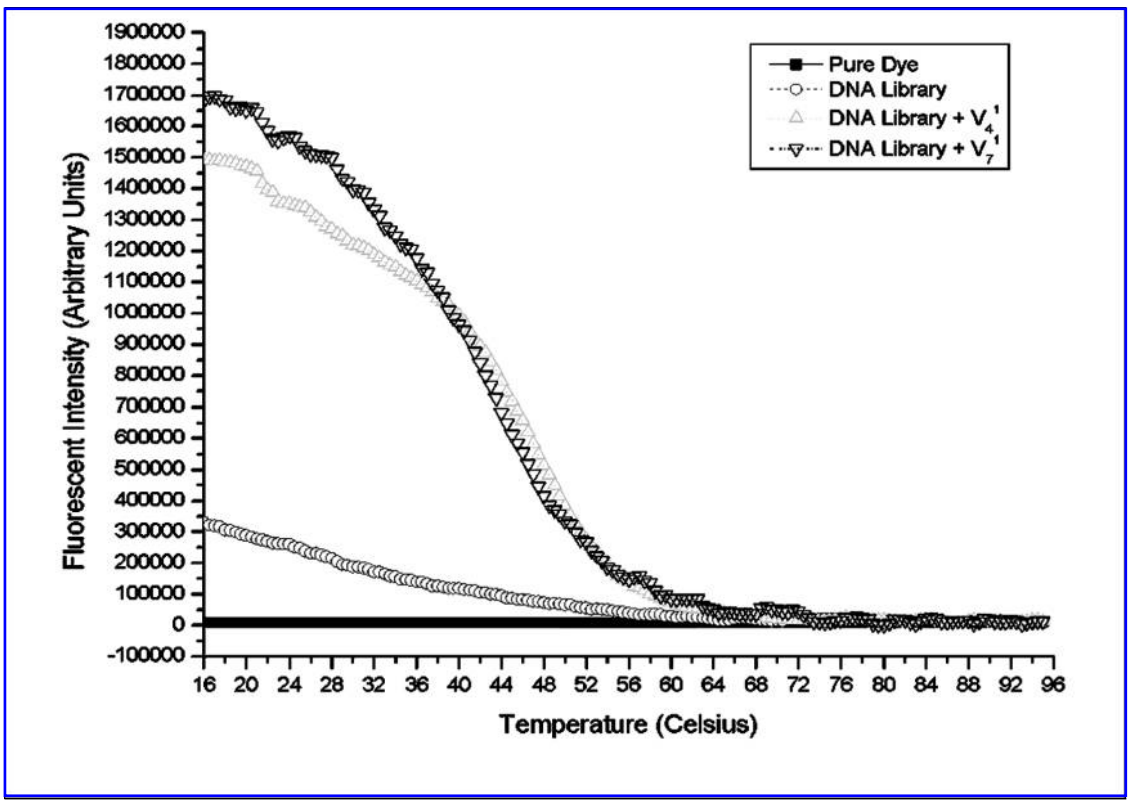

FIG. 7. Testing the library sequences for a hybridization and a secondary structure formation. Melting curves of the whole DNA library (2 μM) are measured in the presence of a capture probe V_4^1 (1 μM) or a capture probe V_7^1 (1 μM) or in the absence of any capture probe in 50 mM NaCl and 10 mM sodium phosphate buffer. The melting points of the capture probes V_4^1 and V_7^1 were found to be 46.3 and 43.7°C, respectively. The fluorescent signal for the DNA library with no capture probe at that temperature is about seven times less than that with capture probes V_4^1 and V_7^1.

DISCUSSION

The approach presented for DNA library design is based mainly on hybridization thermodynamics and depends strongly on the quality of the experimental thermodynamic data available in the literature today. A modeling of the DNA/DNA hybridization kinetics as a quantum mechanical interaction between two biopolymers is a major challenge for the next generation of supercomputers. The general applicability of hybridization thermodynamics has been intensively demonstrated over the last decades of scientific research on nucleic acids. Additional constraints, desired in specific experimental set-ups, can easily be implemented.

It was shown by the PCR experiments that all bits are presented in the DNA library and identical conditions exist for all captures probes of the library, in which a rather uniform amplification of the expected PCR products is observed. No significant amplification of unexpected PCR products was observed with one exception involving capture probe V_7^1. The PCR could be successfully used as a readout tool for the individual library sequences.

DNA hybridization analyses on beads suggested that these unexpected PCR products do not indicate significant problems for the selection procedure with the library. The DNA polymerase is very sensitive to mismatches at the $3'$ primer end. One mismatch in the last three nucleotides on the $3'$ end could prevent the elongation (Pirrung *et al.*, 2000). At the same time, it seems that rather unstable mispriming involving a formation of a short double-stranded $3'$ end could be responsible for unexpected PCR amplification even at high annealing temperatures. Such mispriming at the $3'$ end could be avoided in the future designs.

The melting curves obtained suggest one more time that identical hybridization conditions exist for all capture probes in which a fairly uniform hybridisation yield could be expected. The absence of G's in the library sequences proved to be a successful strategy for avoiding the formation of secondary structures within the library sequences and hybridization among them (Braich *et al.*, 2001, 2002; Faulhammer *et al.*, 2000; Mir, 1999).

The high level of a specific hybridization achieved with the presented twelve-bit library opens the important question of whether the method is scalable to large sets. Recently a 20-variable 3-SAT problem has been solved by the Adleman group (Braich *et al.*, 2000) applying multistep DNA selection. Their 20-bit DNA library (40 words, each 15 nt long) has been optimized based on combinatorial constraints (Braich *et al.*, 2001). According to our thermodynamic criteria, their word set showed reasonable properties: a free energy gap of $\delta G = 4.6$ kcal/mol between specific and nonspecific DNA hybridization and a variation of the melting temperatures in a total range of $\Delta T_m = 8.4^\circ$C (according to the nearest neighbor approximation). Surprisingly, our numerical tests show that increasing the set size is not necessarily connected with a significant reduction of the word quality. A free energy gap of $\delta G = 8.8$ kcal/mol and a temperature variation of $\Delta T_m = 5^\circ$C has been obtained by our algorithm for a set of 128 DNA words (each word 16 nt long) for a 64-bit problem. The 128-word set is available from one of the authors (JA) on request. This is a promising result, which needs an experimental verification.

We believe the discrimination between perfectly matching DNA hybrids and those with mismatches based on hybridization thermodynamics is more accurate than using the Hamming distance or the number of slide mismatches. The application of the partition function takes into account the formation of hairpins and internal or bulge loops and increases the efficiency of the thermodynamic discrimination. The constraint for thermodynamically uniform melting points of the words is more accurate than that based on an identical CG content. Our calculations did show that 16-deoxynucleoties with a 50% CG contend could have differences in melting points, estimated thermodynamically, of $\pm 10^\circ$C.

It is clear that general-purpose algorithms could be executed on DNA-based computers solving a large class of search problems. Further integration between the molecular computing and the DNA-chip and nano-technologies is necessary for achieving a fully automated molecular computation.

The presented library design is very suitable for an integrated molecular computation in micro-flow reactors by multiple DNA selection under isothermal conditions (Penchovsky and McCaskill, 2002) because of the uniformity of melting points of the words. The method is not restricted to the design of libraries for DNA computation and has already been considered for the design of PCR primers, the DNA sequences for the isothermal amplification reactions, and capture probes for DNA chip arrays.

ACKNOWLEDGMENTS

We are grateful to Ivo L. Hofacker (Vienna, Austria) for his technical support with the Vienna RNA folding package, and John Santa Lucia, Jr. (Detroit, USA) for permission to use his DNA energy parameters.

REFERENCES

Adleman, L.M. 1994. Molecular computation of solutions to combinatorial problems. *Science* 266, 1021–1024.

Allawi, H.T., and Santa Lucia, Jr., J. 1997. Thermodynamics and NMR of internal GT mismatches in DNA. *Biochemistry* 36, 10581–10594.

Braich, R.S., Chevlyapov, N., Johnson, C., Rothemund, P.W.K., and Adleman, L.M. 2002. Solution of a 20-variable 3-SAT problem on a DNA computer. *Science* 296, 499–502.

Braich, R.S., Johnson, C., Rothemund, P.W.K., Hwang, D., Chevlyapov, N., and Adleman, L.M. 2001. Solution of a satisfiability problem on a gel-based DNA computer. LNCS 2054, 27–42.

Breslauer, K.J., Frank, R., Blöcker, H., and Marky, L.A. 1986. Predicting DNA duplex stability from the base sequence. *Proc. Natl. Acad. Sci. USA* 83, 3746–3750.

Cox, J.C., Cohen, D.S., and Ellington, A.D. 1999. The complexities of DNA computation. *TIBTECH* 17, 151–154.

Deaton, R., Garzon, M., Murphy, R.C., Rose, J.A., Franceschetti, D.R., and Stevens, Jr., S.E. 1998. Reliability and efficiency of a DNA-based Computation. *Phys. Review Letters* 80, 417–420.

Fan, Z.H., Mangru, S., Granzow, R., Heaney, P., Ho, W., Dong, Q., and Kumar, P. 1999. Dynamic DNA hybridization on a chip using paramagnetic beads. *Anal. Chem.* 71, 4851–4859.

Faulhammer, D., Cukras, A.R., Lipton, R.J., and Landweber, L.F. 2000. Molecular computation: RNA solution to chess problems. *Proc. Natl. Acad. Sci.* 97, 1385–1389.

Frutos, A.G., Liu, Q., Thiel, A.J., Sanner, A.M.W., Condon, A.E., Smith, L.M., and Corn, R.M. 1997. Demonstration of a word design strategy for DNA computing on surfaces. *Nucl. Acids Res.* 25, 4748–4757.

Gifford, D.K. 1994. On the path to computation with DNA. *Science* 266, 993–994.

Hofacker, I.L., Fontana, W., Stadler, P.F., Bonhoeffer, S., Tacker, M., and Schuster, P. 1994. Fast folding and comparision of RNA secondary structures. (The Vienna RNA package), Monatsh. Chem. 125, 167–188.

James, K.D., Boles, A.R., Henckel, D., and Ellington, A.D. 1998. The fidelity of template-directed oligonucleotide ligation and its relevance to DNA computation. *Nucl. Acids Res.* 26, 5203–5211.

Lipton, R.J. 1995. DNA solution of hard computational problems. *Science* 268, 542–545.

Liu, Q., Wang, L., Frutos, A.G., Condon, A.E., Corn, R.M., and Smith, L.M. 2000. DNA computing on surfaces. *Nature* 403, 175–179.

Marathe, A., Condon, A.E., and Corn, R.M. 2001. On combinatorial DNA word design. *J. Comp. Biol.* 8, 201–219.

McCaskill, J.S. 1990. The equilibrium partition function and base pair binding probbilities for RNA secondary structure. *Biopolymers* 29, 1109–1119.

McCaskill, J.S. 2001. Optically programming DNA computing in microflow reactors. *BioSystems* 59, 125–138.

Mir, K.U. 1999. A restricted genetic alphabet for DNA computing. *Proc. DNA Based Computers. DIAMACS Workshop, DIMACS Series in Discrete Mathematics and Theoretical Computer Science* 44, 243–246.

Penchovsky, R., Birch-Hirschfeld, E., and McCaskill, J.S. 2000. End-specific covalent photo-dependent immobilisation of synthetic DNA to paramagnetic beads. *Nucl. Acids Res.* 28, e98.

Penchovsky, R., McCaskill, J.S. 2002. Cascadable hybridisation transfer of specific DNA between microreactor selection modules. *LNCS* 2340, 46–56.

Pevzner, P.A., Tang, H., and Waterman, M.S. 2001. An Eulerian path approach to DNA fragment assembly. *Proc. Natl. Acad. Sci.* 98, 9748–9753.

Pirrung, M.C., Connors, R.V., Odenbaugh, A.L., Montague-Smith, M.P., Nathan, G.W., and Tollett, J.J. 2000. The array primer extension method for DNA microchip analysis. Molecular computation of satisfaction problems. *J. Am. Chem. Soc.* 122, 1873–1882.

Quyang, Q., Kaplan, P.D., Liu, S., and Libchaber, A. 1997. DNA solution of the maximal clique problem. *Science* 278, 446–449.

Sambroock, J., Fritsch, E.F., and Maniatis, T. 1989. Molecular Cloning: A Laboratory Manual. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Santa Lucia, Jr. J. 1998. A unified view of polymer, dumbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc. Natl. Acad. Sci.* 95, 1460–1465.

Stevens, P.W., Henry, M.R., and Keslo, D.M. 1999. DNA hybridisation on microparticles: Determining capture-probe density and equilibrium dissociation constants. *Nucl. Acids Res.* 27, 1719–1727.

Tang, J., and Breaker, R.R. 1998. Mechanism for allosteric inhibition of an ATP-sensitive ribozyme. *Nucl. Acids Res.* 26, 4214–4221.

Wetmure, J.G. 1991. DNA probes: Application of the principle of nucleic acid hybridisation. Critical Reviews in Biochemistry and Molecular Biology 26, 227–259.

Winfree, E., Liu, F., Wenzler, L.A., and Seeman, N.C. 1998. Design and self-assembly of two-dimensional DNA crystals. *Nature* 394, 539–544.

Address correspondence to:
Robert Penchovsky
Biomolecular Information Processing
Fraunhofen Gesellschaft
Schloss Birlinghoven
D-53754 Sankt Augustin
Germany

E-mail: Robert_Penchovsky@web.de

This article has been cited by:

1. Susannah Gal, Nancy Monteith, Anthony J. Macula. 2008. Successful preparation and analysis of a 5-site 2-variable DNA library. *Natural Computing* . [CrossRef]
2. Morgan A. Bishop , Arkadii G. D'Yachkov , Anthony J. Macula , Thomas E. Renz , Vyacheslav V. Rykov . 2007. Free Energy Gap and Statistical Thermodynamic Fidelity of DNA Codes. *Journal of Computational Biology* **14**:8, 1088-1104. [Abstract] [PDF] [PDF Plus]
3. Z. Ibrahim, Y. Tsuboi, O. Ono. 2006. Hybridization-Ligation Versus Parallel Overlap Assembly: An Experimental Comparison of Initial Pool Generation for Direct-Proportional Length-Based DNA Computing. *IEEE Transactions on Nanobioscience* **5**:2, 103-109. [CrossRef]
4. J. Chen, R. Deaton, M. Garzon, J. -W. Kim, D. H. Wood, H. Bi, D. Carpenter, Y. -Z. Wang. 2006. Characterization of Non-crosshybridizing DNA Oligonucleotides Manufactured in vitro. *Natural Computing* **5**:2, 165-181. [CrossRef]
5. Robert Penchovsky, Ronald R Breaker. 2005. Computational design and experimental validation of oligonucleotide-sensing allosteric ribozymes. *Nature Biotechnology* **23**:11, 1424-1433. [CrossRef]
6. S.-Y. Shin, I.-H. Lee, D. Kim, B.-T. Zhang. 2005. Multiobjective Evolutionary Optimization of DNA Sequences for Reliable DNA Computing. *IEEE Transactions on Evolutionary Computation* **9**:2, 143-158. [CrossRef]