DNA Word Design Strategy for Creating Sets of Non-interacting Oligonucleotides for DNA Microarrays

Ming Li,† Hye Jin Lee,† Anne E. Condon,‡ and Robert M. Corn*,†

Department of Chemistry, University of Wisconsin—Madison, 1101 University Avenue, Madison, Wisconsin 53706, and Department of Computer Science, University of British Columbia, 201-2366 Main Mall, Vancouver, British Columbia V6T1Z4, Canada

Received August 1, 2001. In Final Form: October 22, 2001

A template−map design strategy for generating sets of non-interacting DNA oligonucleotides for applications in DNA arrays and biosensors is demonstrated. This strategy is used to create a set of oligonucleotides of size s with length l that possess at least n base mismatches with the complements of all the other members in the set. These "DNA word" sets are denoted as $n$bm l-mers or l:n sets. To regularize the thermodynamic stability of the perfectly matched hybridized DNA duplexes, the l-mers chosen for all the sets are required to have an approximately 50% G/C content. To achieve good discrimination between each DNA word in each set generated using the template−map strategy, it is required that n should be approximately equal to $l/2$ or higher. The template−map strategy can be used in a straightforward manner to create DNA word sets for cases when $l = 4k$ and $n = 2k$, where k is an integer. Specific examples of $4k$:$2k$ sets are designed: an 8:4 set ($s = 224$), a 12:6 set ($s = 528$), a 16:8 set ($s = 960$), and a 20:10 set ($s = 1520$). These sets are further optimized to achieve the narrowest possible distribution of melting temperatures by selecting the best set after permutation of the templates and maps over all possible configurations. To demonstrate the viability of this methodology, a non-interacting set of four specific 6bm 12mers have been chosen, synthesized, and used in an SPR imaging measurement of the hybridization adsorption onto a DNA array. The template−map strategy is also applied to generate DNA word sets for cases where $l \neq 4k$. In these cases, the creation of the maps and templates is more complicated, but possible. The templates and maps for three additional types of sets are created: $(4k - 1)$:$(2k - 1)$, $(4k + 1)$:$2k$, and $(4k - 2)$:$(2k - 1)$. Specific examples are given for $l = 7$, 9, and 10: DNA word sets of 7:3 ($s = 224$), 9:4 ($s = 360$), and 10:5 ($s = 132$).

1. Introduction

There are currently several fields that rely heavily on the hybridization of sets of short (<30 bases) oligonucleotides for both biological and nonbiological applications. These include the hybridization adsorption onto DNA arrays for biosensor applications,[1−3] the creation of biomolecular-based computational systems,[4−6] and the formation of novel nanostructured materials with unique optical and transport properties.[7] The highly predictable hybridization chemistry of DNA, the ability to completely control the length and content of oligonucleotides, and the wealth of enzymes available for modification of DNA make the nucleic acids attractive for all of these applications. In some cases, it is required that the single stranded DNA (ssDNA) molecules in a set only interact with their perfect complements to form double stranded DNA (dsDNA), and not bind to any other complementary oligonucleotides in the solution. Examples of applications for these well-behaved "DNA word" sets of oligonucleotides are the creation of non-interacting DNA tags that can be used in the formation of universal chips similar to the "zip-code arrays" of Affymetrix.[8−11]

Intensive efforts have focused on designing non-interacting DNA words for DNA computing using combinatorial constraints on the composition of a set of DNA code words for specific applications.[12−22] We have recently

* To whom correspondence should be addressed. Telephone: 1 608 262 1562. Fax: 1 608 262 0453. E-mail: corn@chem.wisc.edu.

† University of Wisconsin—Madison.

‡ University of British Columbia.

(1) Lin, V. S.-Y.; Motesharei, K.; Dancil, K.-P. S.; Sailor, M. J.; Ghadiri, M. R. *Science* **1997**, *278*, 840−843.

(2) Chee, M.; Yang, R.; Hubbell, E.; Berno, A.; Huang, X. C.; Stern, D.; Winkler, J.; Lockhart, D. J.; Morris, M. S.; Fodor, S. P. A. *Science* **1996**, *274*, 610−614.

(3) Southern, E. M.; Case-Green, S. C.; Elder, J. K.; Johnson, M.; Mir, K. U.; Wang, L.; Williams, J. C. *Nucleic Acids Res.* **1994**, *22*, 1368−1373.

(4) Frutos, A. G.; Liu, Q.; Thiel, A. J.; Sanner, A. W.; Condon, A. E.; Smith, L. M.; Corn, R. M. *Nucleic Acids Res.* **1997**, *25*, 4748−4757.

(5) Frutos, A. G.; Smith, L. M.; Corn, R. M. *J. Am. Chem. Soc.* **1998**, *120*, 10277−10282.

(6) Liu, Q.; Wang, L.; Frutos, A. G.; Condon, A. E.; Corn, R. M.; Smith, L. M. *Nature* **2000**, *403*, 175−179.

(7) Liu, J.; Zhang, Q.; Remsen, E. E.; Wooley, K. L. *Biomacromolecules* **2001**, *2*, 362−368.

(8) Brenner, S.; Johnson, M.; Bridgham, J.; Golda, G.; Lloyd, D. H.; Johnson, D.; Luo, S.; McCurdy, S.; Foy, M.; Ewan, M.; Roth, R.; George, D.; Eletr, S.; Albrecht, G.; Vermaas, E.; Williams, S. R.; Moon, K.; Burcham, T.; Pallas, M.; DuBridge, R. B.; Kirchner, J.; Fearon, K.; Mao, J.; Corcoran, K. *Nat. Biotechnol.* **2000**, *18* (6), 630−634.

(9) Winzeler, E. A.; Richards, D. R.; Conway, A. R.; Goldstein, A. L.; Kalman, S.; McCullough, M. J.; McCusker, J. H.; Stevens, D. A.; Wodicka, L.; Lockhart, D. J.; Davis, R. W. *Science* **1998**, *281*, 1194−1197.

(10) Schena, M.; Shalon, D.; Davis, R. W.; Brown, P. O. *Science* **1995**, *270*, 467−470.

(11) Pease, A. C.; Solas, D.; Sullivan, E. J.; Cronin, M. T.; Holmes, C. P.; Fodor, S. P. A. *Proc. Natl. Acad. Sci. U.S.A.* **1994**, *91*, 5022−5026.

(12) Braich, R. S.; Johnson, C.; Rothemud, P. W. K.; Hwang, D.; Chelyapov, N.; Adleman, L. M. *Solution of a satisfiability problem on a gel based DNA computer*; The Netherlands: Leiden, June, 2000.

(13) Faulhammer, D.; Cukras, A. R.; Lipton, R. J.; Landweber, L. F. *Proc. Natl. Acad. Sci.* **2000**, *97* (4), 1385−1389.

(14) Sakamoto, K.; Gouzu, H.; Komiya, K.; Kiga, D.; Yokoyama, S.; Yokomori, T.; Hagiya, M. *Science* **2000**, *288*, 1223−1226.

(15) Brenner, S. *Methods for Sorting Polynucleotides using Oligonucleotide Tags.* U.S.A. patent dated Feb 18, 1997, ISBN # 5604097.

(16) Shoemaker, D. D.; Lashkari, D. A.; Morris, D.; Mittman, M.; Davis, R. W. *Nat. Genet.* **1996**, *16*, 450−456.

(17) Mir, K. U. A Restricted Genetic Alphabet for DNA Computing. Proceedings of DNA Based Computers II, DIMACS Workshop, June 10−12, 1996; Landweber, L. F., Baum, E. B., Eds.; *DIMACS Ser. Discrete Math. Theor. Comput. Sci.* **1999**, *44*, 243−246.

(18) Deaton, R.; Garzon, M.; Murphy, R. C.; Rose, J. A.; Franceschetti, D. R.; Stevens, S. E., Jr. In *Genetic Search of Reliable Encodings for DNA-Based Computation*; Koza, J. R., Goldberg, D. E., Fogel, D. B., Riolo, R. L., Eds.; Proceedings of the First Annual Conference on Genetic Programming; 1996.

(19) Baum, E. B. *DNA Sequences Useful for Computation*. Proc. 2nd DIMACS Workshop on DNA Bases Computers; June 1996.

10.1021/la0112209 CCC: $22.00
Published on Web 01/05/2002

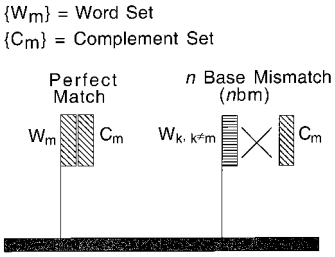

Figure 1. Schematic cartoon showing the DNA computing strategy used to generate a word set $\{W_m\}$ defined also as an *l*:*n* set. For a given word set $\{W_m\}$ and the corresponding complement set $\{C_m\}$, each member W_m of length *l* forms a DNA duplex with its perfectly matched complement, C_m. The pair of those molecules is referred to as the "perfect match", and all the other pairs $W_{k,k\neq m}/C_m$ contain at least *n* base mismatched pairs. This set $\{W_m\}$ is called an *l*:*n* set.

employed the generation of sets of DNA words in the demonstration of DNA computing at surfaces.[6] In this work, a word design strategy involving a set of 16 base oligonucleotides ("16mers") was investigated to store four bits of information for a four-variable satisfiability (SAT) calculation in a demonstration of a prototype DNA computer.[6] Figure 1 shows the DNA word design strategy as described in a previous paper,[5] in which each word in a set can be uniquely distinguished from all other words on the surface by the hybridization of complements. Each member W_m of the word set has at least *n* base locations different from all the other elements. If *n* is large enough, only the "perfect match" pair of molecules W_m/C_m will form a DNA duplex, while all other pairs of words $W_{k,k\neq m}/C_m$ containing at least *n* base mismatched pairs will not hybridize. A set of molecules of length *l* in which all mismatches are greater than or equal to *n* is denoted as a set of *n*bm *l*-mers or an *l*:*n* set.

In this paper, we demonstrate a general template−map strategy for designing *l*:*n* sets of non-interacting oligonucleotides. A template is a nucleotide sequence over two bases, and a map is a string of binary variables. DNA word sequences are generated by the operation of each map in a given set upon each template in another set. The template−map strategy works particularly well for the case of $l = 4k$ and $n = 2k$, where *k* is an integer, and specific examples of 4*k*:2*k* sets are given for when $k = 2$, 3, 4, and 5. DNA word sets can also be generated when $l \neq 4k$, and the template−map strategy is used to create sets of three additional types: $(4k - 1):(2k - 1)$, $(4k + 1):2k$, and $(4k - 2):(2k - 1)$. Different mathematical treatments are required to generate maps for the $4k:2k$, $(4k - 1):(2k - 1)$, $(4k + 1):2k$, and $(4k - 2):(2k - 1)$ sets, and the corresponding specific examples of each set type are given. For all of the sets, the G/C content of every member is fixed at approximately 50%, so that each perfectly matched hybridized DNA duplex (dsDNA) has a similar thermodynamic stability. For the $4k:2k$ ($k = 2$, 3, 4, and 5) sets, the melting temperature (T_m) and standard Gibbs free energy ($\Delta G°$) of hybridization of all the oligonucleotides generated are evaluated using simplified thermodynamics calculations.[23] For the specific case of the set of 6bm 12mers, four different words (W_1−W_4) are chosen to experimentally demonstrate the utility of this template−map design strategy. The stability of these duplexes (W_m−C_m) is analyzed using melting temperature measurements, and as a final check, surface plasmon resonance (SPR) imaging measurements are used to study the specific hybridization adsorption of the corresponding complementary ssDNA molecules (C_1−C_4) onto the surface bound ssDNA words (W_1−W_4).

2. Experimental Considerations

A. Materials. The chemicals 11-mercaptoundecylamine (MUAM) (Dojindo), sulfosuccinimidyl 4-(*N*-maleimidomethyl)-cyclohexane-1-carboxylate (SSMCC) (pierce), urea (Bio-Rad Laboratories), 9-fluorenylmethoxycarbonyl-*N*-hydroxysuccinimide (Fmoc-NHS) (Novabiochem), and triethanolamine hydrochloride (TEA) (Sigma) were all used as received. Gold thin films (45 nm) used for SPR imaging measurements were vapor deposited onto SF-10 substrates (18×18 mm^2, Schott Glass) as reported previously.[24] Millipore filtered water was used for all aqueous solutions and rinsing. All oligonucleotides were synthesized on an ABI DNA synthesizer at the University of Wisconsin Biotechnology Center. Deprotection and purification of oligonucleotides were performed as described previously.[25,26] The buffer used for SPR imaging experiments contained 20 mM phosphate (pH 7.4), 100 mM NaCl, 1 mM EDTA, 1 mM DTT, and 5 mM $MgCl_2$. Removal of hybridized complementary molecules was accomplished by exposing the surface to 8 M urea at room temperature for 15 min.

B. DNA Surface Attachment Chemistry. The covalent attachment of DNA oligonucleotides onto gold thin films has been reported previously.[5,27] Briefly, MUAM was self-assembled from an ethanolic solution onto a gold-coated glass substrate. The MUAM self-assembled monolayer on the gold surface was then reacted with the hydrophobic protecting group, Fmoc-NHS, and the hydrophobic surface was photopatterned to create an array of bare gold areas, followed by the adsorption of MUAM to fill in the bare gold array elements. The amine-terminated gold surface was reacted with the heterobifunctional linker, SSMCC. The thiol-reactive maleimide-terminated surface was then reacted with single-stranded 5′-thiol modified DNA for at least 4 h. The fabrication of the multicompoment DNA arrays for SPR imaging experiments has been shown in a previous paper.[28]

C. Melting Temperature Measurements. DNA melting temperature curves were obtained by monitoring the absorbance of DNA solutions at 260 nm as a function of temperature using an HP8452A UV−vis spectrophotometer equipped with an HP89090A Peltier temperature control accessory. Melting temperatures (T_m's) were measured in buffer solutions (pH 7) containing 10 mM sodium phosphate, 1 mM EDTA, 1 M NaCl, and 2 μM oligonucleotides. A ramp rate of 1 °C/min with a hold time of 1 min was used over the range 25−90 °C to record melting temperature curves of ssDNA molecules. The T_m (if observed) was determined as the temperature at which the first derivative of the raw UV absorbance curve reached the maximum and was estimated within the error ±1.5 °C.

D. SPR Imaging Apparatus. The in situ SPR imaging instrument has been described previously.[25] Briefly, a collimated white light source was used to illuminate a gold film (45 nm)/prism interface at a fixed incident angle near the SPR angle. The reflected light was passed through a 10 nm band-pass filter ($\lambda = 830$ nm) and was collected with an inexpensive CCD camera (GWC Instruments). Differences in the reflected light intensity are a direct result of differences in the refractive index of the material bound at the gold surface. The images shown in this work were analyzed using NIH Image v.1.61 software.

(20) Deaton, R.; Murphy, R. C.; Garzon, M.; Franceschetti, D. R.; Stevens, S. E., Jr. *Good Encoding for DNA-Based Solutions to Combinatorial Problems*. Proc. of DNA Based Computers II, DIMACS Workshop, June 10−12, 1996; Landweber, L. F., Baum, E. B., Eds.; *DIMACS Ser. Discrete Math. Theor. Comput. Sci.* **1999**, *44*, 247−258.
(21) Adleman, L. M. *Science* **1994**, *266*, 1021−1024.
(22) Brenner, S.; Lerner, R. A. *Proc. Natl. Acad. Sci.* **1992**, *89*, 5381−5383.
(23) Liu, Q.; Frutos, A. G.; Thiel, A. J.; Corn, R. M.; Smith, L. M. *J. Comput. Biol.* **1998**, *5*, 269−278.
(24) Hanken, D. G.; Corn, R. M. *Anal. Chem.* **1995**, *67*, 3767−3774.
(25) Nelson, B. P.; Frutos, A. G.; Brockman, J. M.; Corn, R. M. *Anal. Chem.* **1999**, *71*, 3935−3940.
(26) *User Guide to DNA Modification and Labeling*; Glen Research Corporation: Sterling, VA, 1990.
(27) Jordan, C. E.; Frutos, A. G.; Thiel, A. J.; Corn, R. M. *Anal. Chem.* **1997**, *69*, 4939−4947.
(28) Brockman, J. M.; Frutos, A. G.; Corn, R. M. *J. Am. Chem. Soc.* **1999**, *121*, 8044−8051.

Table 1. List of DNA Word Sets for $l = 4-20$

length (l)	$n$bm	set size (s)	type[a]
4	2	48	I
5	2	120	II
6	3	56	IV
7	3	224	III
8	4	224	I
9	4	360	II
10	5	132	IV
11	5	528	III
12	6	528	I
13	6	728	II
14	7	240	IV
15	7	960	III
16	8	960	I
17	8	1224	II
18	9	380	IV
19	9	1520	III
20	10	1520	I

[a] Type I, II, III, and IV sets correspond to $4k$:$2k$, $(4k + 1)$:$2k$, $(4k - 1)$:$(2k - 1)$, and $(4k - 2)$:$(2k - 1)$, respectively.

3. Results and Discussion

Using the template−map strategy, a DNA word set can be created for any length l and mismatch number n. The sizes of word sets that we have created for $l = 4-20$ are listed in Table 1. As seen in the table, particularly large sets can be easily created for the case where $l = 4k$ and $n = 2k$, where $k = 2$, 3, 4, and 5. In the table, we denote these $4k$:$2k$ sets as type I DNA word sets, and we now describe in detail how to create the maps and templates for this case.

A. $4k$:$2k$ Set Generation. The mathematical treatment of generating non-interacting oligonucleotides is based on a template−map strategy[4] which utilizes hamming codes to provide the templates and maps that are needed to generate the set of DNA sequences. Depending on the number of maps and templates, oligonucleotide sets of different sizes can be generated. Each template−map pair (t, m) generates a sequence satisfying the rule that, for each 1-bit in the map, the corresponding bit in the template is changed to its complement and, for each 0-bit, the template remains unchanged. As an example, a template−map pair (ACAACCAA, 01100110) is used to generate the 8mer sequence AGTACGTA as shown in Figure 2. In this case, a total number of 16 maps and 14 templates were used to generate a set of 224 4bm 8mers. Each map in the set has at least a four base position mismatch with all of other maps.

To find a map set, M, together with a template set, T, the hamming code was employed with two given constraints (see the Appendix, section A for details): (i) the G/C content for each set is fixed at approximately 50% to achieve similar thermodynamic stability of each perfectly matched hybridized DNA duplex, and (ii) any two DNA oligonucleotides with length l in the set differ in at least $n(n \approx l{:}2)$ places. Each DNA sequence can be uniquely distinguished from all other sequences by the hybridization of its complement. The hamming code is described as follows: Let $x = x_1x_2...x_n$ be a word and $y = y_1y_2...y_n$ be another word where x and y are over the binary alphabet {0, 1}. The hamming distance $H(x,y)$ is the number of indices, where $x_i \neq y_i$. The hamming constraint, with distance parameter n, is that for all pairs of distinct words, (x, y) in the set, $H(x,y) \geq n$. Hamming binary code, (l, G, n), is a set of G vectors over {0, 1} of length l such that any two vectors differ in at least n places.[29] To generate the maximal number of sequences of $4k$:$2k$ sets, where k is an integer, two conditions are employed: (i) the map set, M, is the same as the hamming code $(l, 2l, l/2)$ using

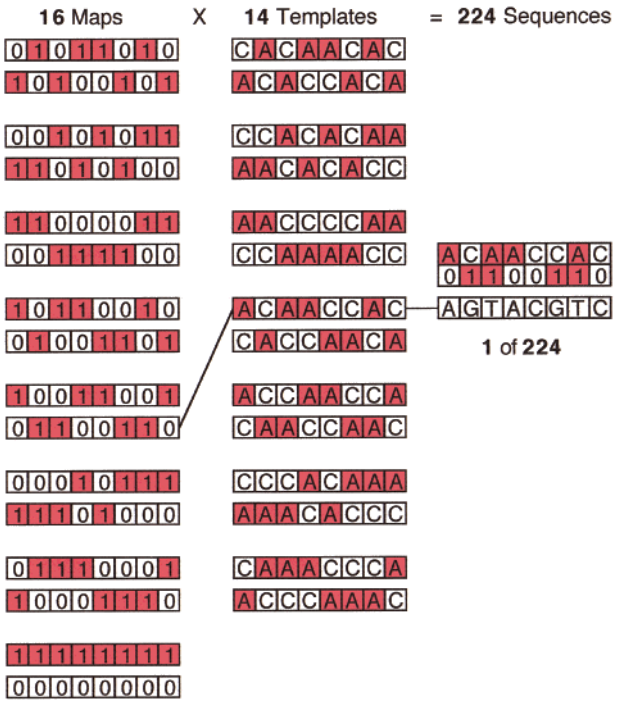

Figure 2. Sets of maps and templates used for generating the l:n = 8:4 set. 8mers are generated by crossing each template with each map using the following rule: for each position in a map with 1, the corresponding position in the template is changed to the complementary base, while, for each position in a map with 0, the corresponding position in the template is unchanged.

Hadamard matrixes (see the Appendix, section B for details),[29] and (ii) the template set, T, is similar to the map set, M, excluding the all-zeros vector and the all-ones vector, and replacing {0, 1} in the hamming code with a different base pair such as {A, C}, {A, G}, {T, C}, or {T, G} and vice versa. Here, {0, 1} was replaced with {C, A} in generating the template set T. This strategy creates a "type I" set of size $s(4k{:}2k)$, that can be expressed as

$$s(4k{:}2k) = 16k(4k - 1), \text{ where } k \text{ is } 1, 2, 3, \ldots \quad (1)$$

For example, s(8:4) is equal to 224, and s(20:10) is equal to 1520. The sizes for the type I sets where $k = 1-5$ are listed in Table 1. Tables 2−4 list all of the maps and templates required to generate the 12:6, 16:8, and 20:10 sets, respectively. A further breakdown of the types of mismatch pairs for these sets is listed in Table 5. Note in this table how only specific types of mismatches appear; this is due to the $4k$:$2k$ symmetry. A more detailed mathematical treatment of how to use the Hamming codes to create these maps and templates is described in the Appendix and is available on our Web site, http://corndog.chem.wisc.edu.

B. Melting Temperature and Standard Gibbs Free Energy Calculation. For applications of DNA microarrays, it is often important to achieve similar standard Gibbs free energies of hybridization and melting temperatures (T_m) for the perfectly matched hybridized DNA duplexes on the surface. For the $4k$:$2k$ sets of oligonucle-

(29) MacWilliams, F. J.; Sloane, N. J. A. In *The theory of error correcting codes*; Artin, H., Bass, H., Eells, J., Feit, W., Freyd, P. J., Gehring, F. W., Halberstam, H., Hormander, L. V., Kac, M., Kemperman, J. H. B., Lauwerier, H. A., Luxemburg, W. A. J., Peterson, F. P., Singer, I. M., Zaanen, A. C., Eds.; North-Holland Publishing Company: Amsterdam, New York, Oxford, 1978; Vol. 16, pp 1−58 and 673−690.

Table 2. Maps and Templates for the 12:6 Set

12mers (total 24*22=528)	
Maps	**Templates**
000000000000	
111111111111	
110111000100	AACAAACCCACC
001000111011	CCACCCAAACAA
011011100010	CAACAAACCCAC
100100011101	ACCACCCAAACA
101101110000	ACAACAAACCCC
010010001111	CACCACCCAAAA
010110111000	CACAACAAACCC
101001000111	ACACCACCCAAA
001011011100	CCACAACAAACC
110100100011	AACACCACCCAA
000101101110	CCCACAACAAAC
111010010001	AAACACCACCCA
100010110110	ACCCACAACAAC
011101001001	CAAACACCACCA
110001011010	AACCCACAACAC
001110100101	CCAAACACCACA
111000101100	AAACCCACAACC
000111010011	CCCAAACACCAA
011100010110	CAAACCCACAAC
100011101001	ACCCAAACACCA
101110001010	ACAAACCCACAC
010001110101	CACCCAAACACA

Table 3. Maps and Templates for the 16:8 Set

16mers (total 32*30=960)	
Maps	**Templates**
0000000000000000	
1111111111111111	
0101010101010101	CACACACACACACACA
1010101010101010	ACACACACACACACAC
0011001100110011	CCAACCAACCAACCAA
1100110011001100	AACCAACCAACCAACC
0110011001100110	CAACCAACCAACCAAC
1001100110011001	ACCAACCAACCAACCA
0100111100001011	CACCAAAACCCCACAA
1011000011110100	ACAACCCCAAAACACC
0001101001011110	CCCAACACCACAAAAC
1110010110100001	AAACCACAACACCCCA
0111110000111000	CAAAAACCCCAAACCC
1000001111000111	ACCCCCAAAACCCAAA
0010100101101101	CCACACCACAACAACA
1101100101100010	AACAACCACAACCCAC
0100000011111011	CACCCCCCAAAAACAA
1011111100000100	ACAAAAAACCCCCACC
0001010110101110	CCCACACAACACAAAC
1110101001010001	AAACACACCACACCCA
0111001111001000	CAAACCAAAACCACCC
1000110000110111	ACCCAACCCCAACAAA
0010011010011101	CCACCAACACCAAACA
1101011010010010	AACACAACACCACCAC
0000111111110000	CCCCAAAAAAAACCCC
1111000000001111	AAAACCCCCCCCAAAA
0101101010100101	CACAACACACACCACA
1010010101011010	ACACCACACACAACAC
0011110011000011	CCAAAACCAACCCCAA
1100001100111100	AACCCCAACCAAAACC
0110100110010110	CAACACCAACCACAAC
1001011001101001	ACCACAACCAACACCA

otides, melting temperatures were calculated for the perfectly matched duplexes using the set of parameters given by Breslauer et al.[30] Melting temperatures were also calculated for mismatched duplexes in these sets using a simple method[23] which modifies the nearest-neighbor pair model of Breslauer et al. by excluding any nearest-neighbor pairs that contain a mismatched base. Figure 3 shows the distribution of melting temperatures for the perfect matched duplexes of the sequences generated in the 12:6, 16:8, and 20:10 sets. As expected, the average melting temperature rises as the oligonucleotide length

Table 4. Maps and Templates for the 20:10 Set

20mers (total 40*38=1520)	
Maps	**Templates**
00000000000000000000	
11111111111111111111	
01011000010101111001	CACAACCCCACACAAAACCA
10100111101010000110	ACACCAAAACACACCCCAAC
01101100001010111100	CAACAACCCCACACAAAACC
10010011110101000011	ACCACCAAAACACACCCCAA
00110110000101011110	CCAACAACCCCACACAAAAC
11001001111010100001	AACCACCAAAACACACCCCA
00011011000010101111	CCCAACAACCCCACACAAAA
11100100111101010000	AAACCACCAAAACACACCCC
01001101100001010111	CACCAACAACCCCACACAAA
10110010011110101000	ACAACCACCAAAACACACCC
01100110110000101011	CAACCAACAACCCCACACAA
10011001001111010100	ACCAACCACCAAAACACACC
01110011011000010101	CAAACCAACAACCCCACACA
10001100100111101010	ACCCAACCACCAAAACACAC
01111001101100001010	CAAAACCAACAACCCCACAC
10000110010011110101	ACCCCAACCACCAAAACACA
00111100110110000101	CCAAAACCAACAACCCCACA
11000011001001111010	AACCCCAACCACCAAAACAC
01011110011011000010	CACAAAACCAACAACCCCAC
10100001100100111101	ACACCCCAACCACCAAAACA
00101111001101100001	CCACAAAACCAACAACCCCA
11010000110010011110	AACACCCCAACCACCAAAAC
01010111100110110000	CACACAAAACCAACAACCCC
10101000011001001111	ACACACCCCAACCACCAAAA
00101011110011011000	CCACACAAAACCAACAACCC
11010100001100100111	AACACACCCCAACCACCAAA
00010101111001101100	CCCACACAAAACCAACAACC
11101010000110010011	AAACACACCCCAACCACCAA
00001010111100110110	CCCCACACAAAACCAACAAC
11110101000011001001	AAAACACACCCCAACCACCA
00000101011110011011	CCCCCACACAAAACCAACAA
11111010100001100100	AAAAACACACCCCAACCACC
01000010101111001101	CACCCCACACAAAACCAACA
10111101010000110010	ACAAAACACACCCCAACCAC
01100001010111100110	CAACCCCACACAAAACCAAC
10011110101000011001	ACCAAAACACACCCCAACCA
00110000101011110011	CCAACCCCACACAAAACCAA
11001111010100001100	AACCAAAACACACCCCAACC

Table 5. Distribution of Mismatch Pairs in W_m/C_k ($m \neq k$) for 8:4, 12:6, 16:8, and 20:10 Sets

length	8mer	12mer	16mer	20mer
*n*bm set	4bm	6bm	8bm	10bm
$\{W_m\}$ set size *s*	224	528	960	1520
4bm	**8512**	0	0	0
5bm	0	0	0	0
6bm	32 256	**23 936**	0	0
7bm	0	0	0	0
8bm	9184	79 200	**82 560**	0
9bm		70 400	0	0
10bm		79 200	0	**117 952**
11bm		0	0	0
12bm		25 520	752 640	49 248
13bm			0	0
14bm			0	787 968
15bm			0	393 984
16bm			85 440	787 968
17bm				0
18bm				49 248
19bm				0
20bm				122 512
total = $s(s-1)$	49 952	278 256	920 640	2 308 880

increases. Note that the melting temperatures for the oligonucleotide sets fall in a relatively narrow range. This indicates that the sequences generated by the word design strategy are similar in a thermodynamic as well as mathematical sense. One of the interesting properties of the template−map strategy is that if any of the two columns in the templates and maps are exchanged simultaneously, a different set of DNA oligonucleotides with the same size can be obtained. This means that, for

(30) Breslauer, K. J.; Frank, R.; Blöcker, H.; Marky, L. A. *Proc. Natl. Acad. Sci. U.S.A.* **1986**, *83*, 3746−3750.

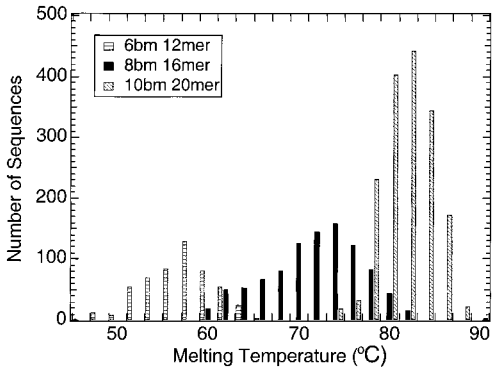

Figure 3. Calculated melting temperature distributions of the 12:6, 16:8, and 20:10 sets using the set of parameters given by Breslauer et al.[18] The oligonucleotide sets in this figure were selected because they show a relatively narrow melting temperature range among all the possible permuted templates and maps listed in Tables 1−3.

Table 6. Thermodynamic Data Including the Melting Temperatures Obtained Experimentally and Theoretically, as Well as Standard Gibbs Free Energies for the 16mers Used in the SPR Imaging Measurements

	12 internal bases[a]	T_m(expt) (°C)	T_m(calc) (°C)	$-\Delta G°$(calc) (kcal/mol)
W_1	CTATGCGTGAAC	68.1	66.2	22.0
W_2	GTATCCGACATG	65.4	66.4	21.6
W_3	GTTAGCCTCAAG	66.4	65.1	21.9
W_4	CATTGCGACTAG	66.0	66.2	22.0

[a] All the words (W_1−W_4) have the sequence 5′-GT*xxxxx-xxxxxxx*TG-3′.

an *l*-mer, the template−map strategy creates [$l(l-1)+1$] different word sets. These DNA word sets are all *n*bm sets, but the melting temperature distribution will differ for each set, which results in a variety of melting temperature distributions. For the 4*k*:2*k* sets, the average melting temperature and standard deviation (σ) were evaluated for all possible configurations after the permutation of the templates and maps. In the case of shorter oligonucleotides (i.e., 8mers), the difference between the conformation with the narrowest distribution range and the one with the widest distribution was larger than the difference for longer oligonucleotides (i.e., 20mers). The final 4*k*:2*k* set was chosen as the configuration that yielded the narrowest possible distribution of melting temperatures. The map and template sets shown in Figure 3 are these selected configurations, and correspond to the templates and maps listed in Tables 2−4. The best σ values for the 8:4, 12:6, 16:8, and 20:10 sets were 6.2, 3.9, 5.7, and 2.7 °C, respectively.

To experimentally verify the uniform stability of duplexes generated by our template−map strategy, we selected four words from the 6bm 12mer set and used them to create four ssDNA molecules of the format 5′-GT*xxxxxxxxxxxx*TG-3′ (where *x* is the 12mer) that we denote as W_1−W_4. Table 6 lists the four sequences along with the calculated and measured solution melting temperatures, T_m. Melting curves were measured for each of the duplexes formed between C_1−C_4 and the words W_1−W_4 in 1 M NaCl with a DNA concentration of 2 μM. As listed in Table 5, a T_m of 65.4−68.1 °C was measured for the perfectly matched duplex W_m−C_m. No melting temperatures were observed for any of the mismatched duplexes, indicating that they are below the starting temperature of the melting curve experiment (25 °C). This result demonstrates that there will be a high degree of discrimination between matched and mismatched duplex pairs, even though these are now 6bm 16mers and not 6bm 12mers. Table 6 also compares $\Delta G°$ and T_m values obtained for each of the four duplexes from both the experiments and using the simple estimation method, described previously.[30] Note that the value from the calculation is in good agreement with the experimentally observed melting temperature of the perfectly matched duplexes, W_m−C_m, despite the fact that the calculated T_m did not take into account the formation of hairpins or other secondary structures.

C. SPR Imaging Measurements. To justify the potential ability of the word sets generated by the word design strategy to be used in DNA microarrays for biosensor applications, SPR imaging measurements were performed on a DNA array composed of the set of oligonucleotides (W_1−W_4) used above. SPR imaging is a surface sensitive technique that can be used to monitor the hybridization adsorption of unlabeled DNA target molecules onto a DNA array attached to a gold thin film that is in optical contact with a prism. The hybridization of complementary DNA onto a surface bound DNA array is indicated by a change in the reflectivity of light from a gold film/prism interface near the SPR angle. We constructed a DNA array containing the four words W_1−W_4 and then monitored the sequential hybridization adsorption of each complement C_1−C_4 to the probe DNA array on the surface. After exposure to one complement, the dsDNA was denatured by rinsing with a solution of 8 M urea.

Figure 4a shows the pattern of four different DNA probes, denoted as W_1−W_4, attached onto a modified gold surface. Figure 4b−e shows the results of four successive hybridizations of C_1−C_4 onto the surface bound W_1−W_4 DNA arrays. Each image shows the difference between two images collected before and after exposure of the surface to one of the complements in the presence of buffer. Note that hybridization is observed only for the perfectly matched spots and shows excellent specificity. The nearly equal SPR signal obtained using W_1−W_4 probe DNA indicates that all the probes have nearly identical accessibility to hybridize with the target molecules from solution. These results confirm that the word design strategy can be employed to design sets of non-interacting oligonucleotides for DNA microarrays used in biosensor applications.

D. $l \neq 4k$-mer Generation. As mentioned previously, the template−map strategy can be used to generate sets of oligonucleotides of any size *l*, but when $l \neq 4k$, the creation of the maps becomes slightly more complicated. In this section, we employ the template−map strategy to generate three additional types of word sets for which *n* is approximately equal to *l*/2. We denoted the 4*k*:2*k* set as type I; we also have generated sets of types II−IV which correspond to $(4k+1):2k$, $(4k-1):(2k-1)$, and $(4k-2):(2k-1)$ sets. Table 1 lists the sizes of the *l*-mer sets that we have created for $l = 4-20$ and also states which type of word set was used. For the "type II" sets, where $l = 4k+1$ and $n = 2k$, it is observed that (i) the map set, M, is the same as the hamming code $(4k+1, 8k+4, 2k)$ using conference matrixes (see the Appendix, section C)[29] and (ii) the template set, T, is similar to the map set, M, excluding the all-zeros vector and the all-ones vector, and replacing {0, 1} in the hamming code with {C, A}. The size of the type II set, denoted as $s[(4k+1):2k]$, is thus given as

$$s[(4k+1):2k] = 8(2k+1)(4k+1), \text{ where } k \text{ is } 1, 2, 3, \ldots \quad (2)$$

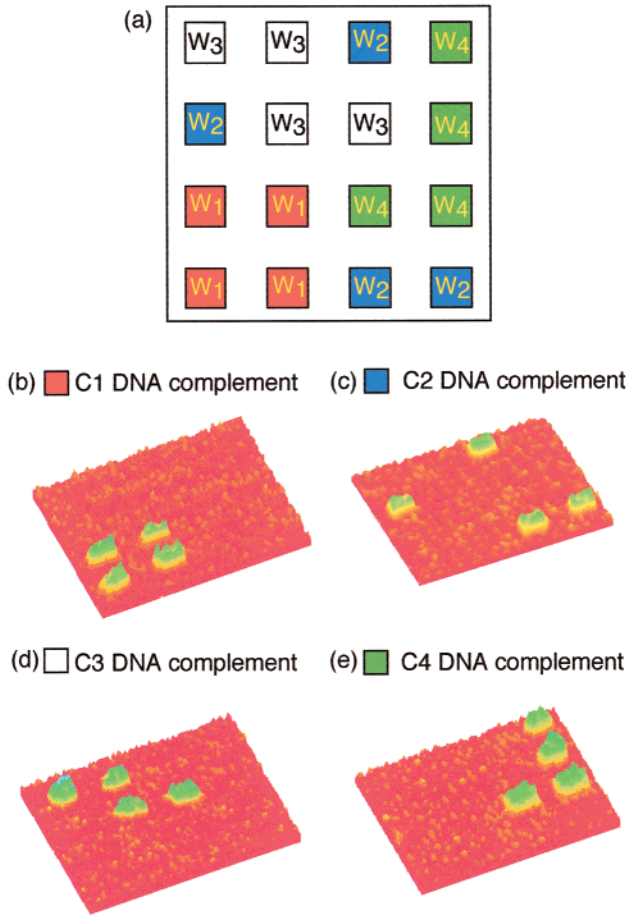

Figure 4. In situ SPR difference images showing the hybridization−adsorption of the complementary DNA (C_1−C_4) onto four different DNA words (W_1−W_4). (a) A schematic diagram showing the pattern of four different probes immobilized on the gold surface. Parts (b)−(e) represent the SPR difference images taken after sequentially hybridizing C_1 through C_4. The hybridization of the complementary DNA (C_1−C_4) onto the surface bound probe DNA array (W_1−W_4) is indicated by a change in the percent reflectivity. The concentration of complementary DNA samples was 100 nM. Between each hybridization, the surface was denatured with 8 M urea. The difference image was obtained by subtracting two images collected immediately before and after exposing the surface to each complement.

For the "type III" word sets where $l = 4k - 1$ and $n = 2k - 1$, it is found that (i) the map set, M, is the same as the hamming code ($4k - 1$, 8k, $2k - 1$) using Hadamard matrixes (see the Appendix, section B)[29] and (ii) the template set, T, is similar to the map set, M, excluding the all-zeros vector and the all-ones vector, and replacing {0, 1} in the hamming code with {C, A}. The size of the type III set, denoted as $s[(4k - 1):(2k - 1)]$, is given as

$$s[(4k - 1):(2k - 1)] = 16k(4k - 1), \text{ where } k \text{ is } 1, 2, 3, \ldots \quad (3)$$

Finally, for the "type IV" word sets where $l = 4k - 2$ and $n = 2k - 1$, (i) the map set, M, is the same as the hamming code ($4k - 2$, $4k$, $2k - 1$), which is achieved from the map set, M′, in $(4k - 1):(2k - 1)$ by taking a cross-section (see the Appendix, section D)[29] and (ii) the template set, T, is similar to the map set, M, excluding the all-zeros vector (notice that the all-ones vector is now not in the M), and replacing {0, 1} in the hamming code with {C, A}. The size of the type IV set, denoted as $s[(4k - 2):(2k - 1)]$ is thus given as

$$s[(4k - 2):(2k - 1)] = 4k(4k - 1), \text{ where } k \text{ is } 1, 2, 3, \ldots \quad (4)$$

Table 7. Distribution of Mismatch Pairs in W_m/C_k ($m \neq k$) for 7:3, 9:4, 10:5, and 17:8 Sets

length	7mer	9mer	10mer	17mer
*n*bm set	3bm	4bm	5bm	8bm
$\{W_m\}$ set size *s*	224	360	132	1224
3bm	**4256**	0	0	0
4bm	4256	**9576**	0	0
5bm	24 192	8568	**1632**	0
6bm	8064	36 720	1360	0
7bm	9184	38 160	4800	0
8bm		26 280	5400	**61 064**
9bm		9936	3800	3400
10bm			300	50 048
11bm				149 056
12bm				394 944
13bm				391 136
14bm				277 984
15bm				69 088
16bm				55 352
17bm				44 880
total = $s(s - 1)$	49 952	129 240	17 292	1 496 952

Sets of oligonucleotides for all lengths $l = 4$−20 have been created using the template−map strategy, and the sizes of these sets are listed in Table 1. A complete listing of the maps and templates, as well as a more detailed description of the creation of these maps and templates for the specific cases of 7:3, 9:4, 10:5, and 17:8 sets, is available on our Web site, http://corndog.chem.wisc.edu. A breakdown of the numbers of different types of mismatch pairs present in these specific $l \neq 4k$-mer sets is summarized in Table 7. Note the lack of symmetry in this table as compared to the case of $4k$:$2k$ sets (Table 5). It should be pointed out that the $(4k - 2):(2k - 1)$ set generates the least number of oligonucleotides compared to that of other cases (i.e. $4k$:$2k$, $(4k - 1):(2k - 1)$, or $(4k + 1):2k$).

4. Conclusions

In this paper we have described a general template−map strategy for designing sets of non-interacting oligonucleotides which can be used for applications in DNA arrays and biosensors. This strategy allows us to generate DNA word sets of any desired length l that have at least an n base mismatch with all other complements in the set, where n is approximately equal to $l/2$. To create these $n$bm l-mers, we employed four types of DNA template−map strategies to create DNA word sets of the form $4k$:$2k$, $(4k - 1):(2k - 1)$, $(4k + 1):2k$, and $(4k - 2):(2k - 1)$. A further selection of the $4k$:$2k$ word sets was made on the basis of the melting temperatures of the various possible template−map configurations. Melting temperature and SPR imaging measurements were used to test four words of the 12:6 set, which confirmed the utility of the template−map strategy to create non-interacting sets of oligonucleotides. These calculations do not yet include any potential hairpin or stem-loop structures; screening for known problem sequences will be the next step in refining these DNA word sets. Compared to this template−map strategy, a random search method can only provide a much smaller set of DNA sequences and is more time-consuming.

5. Appendix: Mathematical Procedure Employed in the Template−Map Strategy for DNA Word Design

The mathematical procedure using a template−map strategy,[4] with the hamming code that finds templates

and maps employed in generating non-interacting oligonucleotides, is shown below. These mathematical treatments were adapted from MacWilliams et al.:[29]

A. Constraints. (a) The G/C content for each set is fixed at approximately 50% to achieve similar thermodynamic stability of each perfectly matched hybridized DNA duplex.

(b) A l:n set is a word set of DNA oligonucleotides of length l such that any two words differ in at least n places. Each DNA sequence can be uniquely distinguished from all other sequences by the hybridization of its complement when $n \approx l/2$.

B. Hadamard Matrixes and Hadamard Codes.[29] (a) A *Hadamard matrix* $\mathbf{H}$ of order l is an $l \times l$ matrix with entries of only +1 and −1 such that

$$\mathbf{H}\mathbf{H}^{\mathrm{T}} = l\mathbf{I}$$

where $\mathbf{I}$ is the identity matrix. For example,

$$\mathbf{H}_1 = (1)$$

$$\mathbf{H}_2 = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$$

$$\mathbf{H}_4 = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix}$$

(b) If $\mathbf{H}_l$ is a *Hadamard matrix* of order l, the following matrix $\mathbf{H}_{2l}$ is then a Hadamard matrix of order $2l$.

$$\mathbf{H}_{2l} = \begin{pmatrix} \mathbf{H}_l & \mathbf{H}_l \\ \mathbf{H}_l & -\mathbf{H}_l \end{pmatrix}$$

Therefore, *Hadamard matrixes* of all orders that are powers of 2 can be obtained.

(c) *Paley Construction*. To obtain a *Hadamard matrix* of order l that is a multiple of 4 but not a power of 2, we can use the Paley construction.

(i) *Quadratic Residues*. Let p be an odd prime. The set Z = {0, 1, ..., p-1} is the set of all possible answers to (x mode p) where x is any non-negative integer. The quadratic residues mod p is a subset of Z that is a set of all squares of non-zero integers mod p. To find this subset, only the following squares need to be considered:

$$1^2, 2^2, ..., [(p-1)/2]^2 \pmod{p}$$

Therefore, the set of quadratic residues mod p contains $(p-1)/2$ values. The remaining numbers in the set Z except zero are called non-residues.

For example, if $p = 13$ the set Z is {0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12}. The quadratic residues mod 13 are

$$(1^2 \bmod 13) = 1, \quad (2^2 \bmod 13) = 4, \quad (3^2 \bmod 13) = 9,$$
$$(4^2 \bmod 13) = 3, (5^2 \bmod 13) = 12, \text{ and}$$
$$(6^2 \bmod 13) = 10$$

The remaining numbers 2, 5, 6, 7, 8 and 11 are the non-residues.

(ii) *Jacobsthal Matrix* $\mathbf{Q} = (q_{ij})$ where $q_{ij} = x(i-j)$ when $i \geq j$

$x(i-j) = 0$ if $(i-j)$ is a multiple of p

$x(i-j) = 1$ if $(i-j)$ mod p is a quadratic residue mod p

$x(i-j) = -1$ if $(i-j)$ mod p is a non-residue

If $i < j$, $q_{ij} = -q_{ji}$

Thus, $\mathbf{Q}$ is skew-symmetric, that is, $\mathbf{Q}^{\mathrm{T}} = -\mathbf{Q}$.

The following Hadamard matrix constructed is of order $l = p + 1$ that is a multiple of 4.

$$\mathbf{H} = \begin{pmatrix} 1 & 1^{\mathrm{T}} \\ 1 & \mathbf{Q} - \mathbf{I} \end{pmatrix}$$

where 1 is a column vector with entries of only +1. From the constructions of sections (a)−(c) together described above, Hadamard matrixes of all orders 1, 2, 4, 8, 12, 16 and 20 can be obtained.

(d) *Hadamard Codes*. Let $\mathbf{H}_l$ be a Hadamard matrix of order l. If +1's are replaced by 0's and −1's by 1's, $\mathbf{H}_l$ is changed into the *binary Hadamard matrix* $\mathbf{A}_l$.

Three Hadamard codes can be obtained from $\mathbf{A}_l$:

(1) ($l-1$, l, $l/2$) code, consisting of the rows of $\mathbf{A}_l$ with the first column deleted;

(2) ($l-1$, $2l$, $l/2-1$) code, consisting of ($l-1$, l, $l/2$) code together with the complements of all its code words;

(3) (l, $2l$, $l/2$) code, consisting of the rows of $\mathbf{A}_l$ and their complements.

C. Conference Matrix and Conference Code.[29] (a) A conference matrix $\mathbf{C}$ of order l is an $l \times l$ matrix with the diagonal entries of 0 and other entries of +1 or −1 such that

$$\mathbf{C}\mathbf{C}^{\mathrm{T}} = (l-1)\mathbf{I}$$

(b) Let $l = p^{\mathrm{m}} + 1$, where p is an odd prime and m is a non-negative integer. Only those l that (l mod 4) = 2 will be considered. We define the matrix $\mathbf{S}$ such that $\mathbf{S} = (q_{ij})$, where $q_{ij} = x(i-j)$ when $i \geq j$, $q_{ij} = q_{ji}$ when $i < j$. Thus, $\mathbf{S}$ is a symmetric square matrix of order $l-1$ satisfying the following equations:

$$\mathbf{S}\mathbf{S}^{\mathrm{T}} = (l-1)\mathbf{I} - \mathbf{J}$$

and

$$\mathbf{S}\mathbf{J} = \mathbf{J}\mathbf{S} = \mathbf{0}$$

where $\mathbf{J}$ is the matrix with all the entries of 1, and $\mathbf{0}$ is the null matrix with all the entries of 0. It is found that

$$\mathbf{C} = \begin{pmatrix} 0 & 1^{\mathrm{T}} \\ 1 & \mathbf{S} \end{pmatrix}$$

(c) *Codes from Conference Matrixes*. The rows of $(\mathbf{S} + \mathbf{I} + \mathbf{J})/2$ and $(-\mathbf{S} + \mathbf{I} + \mathbf{J})/2$ plus the all-zeros and all-ones vectors form an ($l-1$, $2l$, $l/2-1$) conference matrix code.

D. Shortening a Code by Taking a Cross Section. The usual way to shorten a code is to take the code words that begin with 0 and delete that coordinate. For example,(3, 4, 2) can be obtained from (4, 8, 2) by taking a cross section.

$$\begin{pmatrix} 0000 \\ 1111 \\ 0101 \\ 1010 \\ 0011 \\ 1100 \\ 0110 \\ 1001 \end{pmatrix} \rightarrow \begin{pmatrix} 000 \\ 101 \\ 011 \\ 110 \end{pmatrix}$$

Supporting Information Available: Mathematical procedure using a template−map strategy, with the hamming code that finds templates and maps employed in generating non-interacting oligonucleotides. This material is available free of charge via the Internet at http://pubs.acs.org.

LA0112209