Design and test of non-crosshybridizing oligonucleotide building blocks for DNA computers and nanostructures

Russell Deaton,\textsuperscript{a} Jin-Woo Kim,\textsuperscript{b} and Junghuei Chen\textsuperscript{c}

Abstract

DNA oligonucleotides that anneal to form duplexes in specific, planned configurations are a basic construction material for DNA-based computers and nanotechnology. Unplanned duplex configurations introduce errors in computations and defects in structures, and thus, the sequences must be designed to minimize these effects. A software design tool has been developed that uses thermodynamic models of DNA duplex thermal stability and algorithms from graph theory to select good sets of oligonucleotides. An example set was tested in the laboratory, and the designed sequences formed no unplanned duplexes and had no detectable secondary structure.

PACS numbers: 87.14.Gg, 87.15.Aa, 87.15.Cc
DNA has several properties that make it attractive as a construction material for computers and structures on a nanometer scale. With the tools of molecular biology, DNA is easily manipulated in the test tube, can be produced in great quantities of specified size and sequence, and is relatively stable and long-lasting. Most importantly, the reaction in which short, single-stranded DNA sequences (oligonucleotides) anneal to form double-stranded duplexes can be used to program the computation or direct the self-assembly of the nanostructure. For example, in the Adleman-Lipton paradigm of DNA computing (DNAC)\textsuperscript{1–3}, hybridizations between DNA oligonucleotides, which encode the problem, produce a random, exhaustive search of the problem space. In DNA tiling\textsuperscript{4}, hybridizations direct the assembly.

Ideally, the annealing, or hybridization, reaction proceeds by matching the template of one single-stranded DNA with its Watson-Crick complement. The Watson-Crick complements are adenine (A) and thymine (T), and cytosine (C) and guanine (G), respectively. In reality, however, hybridizations can occur with mismatches or non-Watson-Crick base pairs, and between oligonucleotides that are shifted relative to the designed hybridization frame. Hybridizations between oligonucleotides that are not planned are, henceforth, called cross-hybridizations. They result in errors and inefficiency in both computations and structure assembly. Therefore, the DNA sequences that are the building blocks for DNA-based computers and nanostructures have to be designed to minimize cross-hybridizations. These non-cross-hybridizing sequences are also useful for biotechnology applications, like universal DNA microarrays\textsuperscript{5}. This problem, which has been termed the DNA word design problem, is a difficult problem in its own right\textsuperscript{6–8}.

In this paper, a methodology for designing non-crosshybridizing DNA sequences is summarized. This method is detailed elsewhere\textsuperscript{8}. Here, the focus is on experimental results confirming the non-crosshybridizing properties of a test set of oligonucleotides. It was found that the designed sequences did not crosshybridize, and had little to no secondary structure.

The problem of finding a non-crosshybridizing set of DNA oligonucleotides is shown to be equivalent to the problem of finding an independent set of vertices in a graph\textsuperscript{8}, which is NP-complete\textsuperscript{9}. This equivalence suggested that an algorithm for finding maximal independent sets\textsuperscript{10}, (\textit{i. e.} not properly contained in any independent set), as opposed to maximum (\textit{i. e.} maximal independent set of largest size) might be adapted to finding non-crosshybridizing sets of oligonucleotides. The algorithm is:
begin
    $T' \leftarrow 0$
    for $i = 1$ to $m$ do if $i \not\in N(T')$ then $T' \leftarrow T' \cup \{i\}$
end

where $T'$ is the non-crosshybridizing set, and $N(T')$ are all those oligonucleotides that hybridize with a given member of $T'$. In the implementation, an oligonucleotide and its Watson-Crick complement are chosen from an initial starting, random set. Then, all those oligonucleotides that hybridize with the chosen one or its complement are eliminated from consideration. Watson-Crick complements are considered together, as the goal is to produce a set of non-crosshybridizing Watson-Crick complementary pairs. Subsequently, another available oligonucleotide is chosen, and the process is repeated.

An all-or-nothing model of hybridization is assumed. The nearest-neighbor model of duplex thermal stability\textsuperscript{11} is used to determine the energies. Two oligonucleotides are considered to hybridize if their minimum free energies of formation are less than a user-designated threshold. The unified set of thermodynamic parameters from SantaLucia\textsuperscript{11}, as well as parameters for certain mismatches\textsuperscript{12–15}, are used to calculate the total free energy of hybridization ($\Delta G^\circ$) according to the formula,

\[
\Delta G^\circ = \sum_i n_i \Delta G_i^\circ + \Delta G_{GC}(init) + \Delta G_{AT}(init) - 0.114N \ln[Na^+] \tag{1}
\]

where $\Delta G_i^\circ$ is the standard free energy change for the nearest-neighbor pairs in the existing set of data, $n_i$ is the number of nearest-neighbors $i$, and $\Delta G_{GC}(init)$ and $\Delta G_{AT}(init)$ are initiation terms for $G \cdot C$ and $A \cdot T$, respectively. A correction term for self-complementary duplexes is omitted from Eq. 1. The symmetry term is not included in the calculation since the chance of random oligonucleotides being self-complementary is small. In addition, a salt correction term\textsuperscript{11} is applied where $N$ is the total number of phosphates in the duplex divided by 2 (approximately the length), and $[Na^+]$ is the salt concentration.

The minimum free energy of hybridization is computed using a variant of the Smith-Waterman dynamic programming algorithm\textsuperscript{16} for finding local alignments. The scoring
function to construct the matrix of energy values is
\[
\Delta G^o[i][j] = \min \left\{ \begin{array}{l}
\Delta G^o[i][j - 1] + g \\
\Delta G^o[i - 1][j] + g \\
\Delta G^o[i - 1][j - 1] + \Delta G^o_{ij} \\
0
\end{array} \right.,
\]
where $\Delta G^o[i][j]$ is the value of the free energy for the current duplex, $g$ is a penalty applied to dangling ends, loops, bulges, and mismatches not in the parameter set, and $\Delta G^o_{ij}$ is the value of the current nearest-neighbor $ij$. The minimum energy value in the matrix is identified, which corresponds to the minimum free energy of hybridization over the local alignments of the two oligonucleotides.

Using this method, a set of 40 non-crosshybridizing oligonucleotides of length 20 base pairs (bp) were designed. The sequences are given in Table I. The design conditions were a DNA concentration of $1 \times 10^{-6} M$, 1 M salt concentration, reaction temperature of 298.13 K, a gap penalty of 2 kcal/mol, and a threshold of -5 kcal/mol. The average bond energy between these sequences and their complements was -2.64 kcal/mol. In addition, two sequences with predicted secondary structure were selected. They were \textit{ggggggggaaaaacccccccc}, which has a hairpin, as well as secondary structure from more than four consecutive guanines, and \textit{atgcatgcaaaagcatgcat}, which has a hairpin.

Oligonucleotides were synthesized by Integrated DNA Technologies. In addition, the Watson-Crick complement of one of the sequences (#5) was also ordered to serve as a control. All oligonucleotides were purified through a standard desalting process by the supplier. A single band was shown for each oligonucleotide in denaturing TBE-urea polyacrylamide gels (15%), indicating no further purification was necessary. To test the non-crosshybridizing properties of the designed oligonucleotides, they were allowed to anneal for 5 minutes at room temperature. Non-crosshybridization and secondary structure were measured by running the annealing products on TAE polyacrylamide gels (20%) containing 10 mM Mg$^{2+}$ at 4°C. In addition, two sequences mentioned above with planned secondary structure were also ordered, and run in the gels at the same time as the other oligonucleotides to form a basis for comparison.

The results of the experiment are shown in Figures 1 and 2. In Figure 1, lane 1 contains the molecular size marker where the bands indicate duplexes of size 20 to 120 bp in 20 bp increments. Lane 2 is the Watson-Crick complementary pair of sequence 5 (Table I),
lane 3 are the 40 oligonucleotides after annealing, and lane 4 are the 40 oligonucleotides plus the complement of sequence 5. When the 40 oligonucleotides were run together in one lane of the gel after the annealing reaction (lane 3), no duplexes were detected, indicating the non-crosshybridization property. In lanes 5 and 6 are the single oligonucleotides #13 and #23. In lanes 7 and 8 are two single oligonucleotides,\textit{gaggaggggaaaacccccccc} and \textit{atgcatgcaaaagcatgcat}, respectively, with known secondary structure. By comparing the well-defined bands for oligonucleotides 13 and 23 to the smears produced by the oligonucleotides with secondary structure, it was indicated that the designed sequences had little to no secondary structure. In addition, each oligonucleotide was run by itself in a lane of a gel (lanes 1-40) (Figure 2). All of the single oligonucleotides migrated through the gel at a comparable rate indicating that on the time scale of the electrophoresis, no secondary structures were formed.

By capitalizing on the equivalence to the problem of finding an independent set of vertices in a graph, a greedy algorithm was implemented that used the nearest-neighbor model of DNA duplex thermal stability to determine a set of DNA sequences that do not cross-hybridize with each other. A set of oligonucleotides was designed and tested for cross-hybridization. It was found that the designed sequences did not cross-hybridize, and had no detectable secondary structure. Thus, the software design tool should be useful for designing DNA sequences for a variety of biological and other applications, particularly for designing the DNA oligonucleotide building blocks for DNA-based computers and nanotechnology.

This work was supported by NSF Award EIA-0130385. We also thank our collaborators, Dr. David Wood of the University of Delaware, Dr. Max Garzon of the University of Memphis, and Dr. Harvey Rubin of the University of Pennsylvania, for their comments and contributions, and the anonymous reviewer for his comments.
a Computer Science and Engineering, University of Arkansas, Fayetteville, Arkansas, 72701; Electronic address: rdeaton@uark.edu

b Biological and Agricultural Engineering, University of Arkansas, Fayetteville, Arkansas, 72701; Electronic address: jwkim@uark.edu

c Chemistry and Biochemistry, University of Delaware, Newark, Delaware, 19716; Electronic address: junghuei@udel.edu


TABLE I: Non-crosshybridizing library of 40 Watson-Crick pairs. Only one sequence of the pair is shown. The simulation conditions were 23°C, 1 M NaCl, and 1 µM DNA concentration.

<table>
<thead>
<tr>
<th>#</th>
<th>Sequence</th>
<th>#</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>aacaatcttttaagcctaac</td>
<td>2</td>
<td>ttttctatctctgtgat</td>
</tr>
<tr>
<td>3</td>
<td>tgttagagtAAgtggcgg</td>
<td>4</td>
<td>tacccgtgtaaatgtctac</td>
</tr>
<tr>
<td>5</td>
<td>tgtctcaacgattaccccg</td>
<td>6</td>
<td>taccgtaacgtaagccaggg</td>
</tr>
<tr>
<td>7</td>
<td>tccaatgatatgtttaa8</td>
<td>8</td>
<td>tttggatcttatctctgaca</td>
</tr>
<tr>
<td>9</td>
<td>atgaccccaattaggtgtatag</td>
<td>10</td>
<td>gactctatatabttaagacac</td>
</tr>
<tr>
<td>11</td>
<td>ggtccagaaataacagaatcg</td>
<td>12</td>
<td>aacgaacctttctatagtagtac</td>
</tr>
<tr>
<td>13</td>
<td>cagcgtctcctttaagtac</td>
<td>14</td>
<td>gcacaattagcactaaacc</td>
</tr>
<tr>
<td>15</td>
<td>cttccccagtagaattataaa</td>
<td>16</td>
<td>ggaccctgtataacatacaca</td>
</tr>
<tr>
<td>17</td>
<td>gttggaatccactctatatag</td>
<td>18</td>
<td>ctaaaaaagtaataagta</td>
</tr>
<tr>
<td>19</td>
<td>gtttttgtatatgttagctg</td>
<td>20</td>
<td>atctagtgtgttaaatataa</td>
</tr>
<tr>
<td>21</td>
<td>agacttttaggatacactttc</td>
<td>22</td>
<td>atttaagactctctttag</td>
</tr>
<tr>
<td>23</td>
<td>ctttttttctctgtatgcctc</td>
<td>24</td>
<td>ctaacttttgtaagtaaatat</td>
</tr>
<tr>
<td>25</td>
<td>acatcctcataccacactaacat</td>
<td>26</td>
<td>agtaacttcaaccaagcag</td>
</tr>
<tr>
<td>27</td>
<td>ttattactattatacactac</td>
<td>28</td>
<td>gtaattactttcatctaaaa</td>
</tr>
<tr>
<td>29</td>
<td>actagccaaagaatattagta</td>
<td>30</td>
<td>ggtctctgtactttctgact</td>
</tr>
<tr>
<td>31</td>
<td>ttttctaatgtgttatatat</td>
<td>32</td>
<td>aggtttaattatgttcaaatag</td>
</tr>
<tr>
<td>33</td>
<td>tatgcctaggaattataaag</td>
<td>34</td>
<td>cttctctatataatatttc</td>
</tr>
<tr>
<td>35</td>
<td>cctaaagaactcttatatt</td>
<td>36</td>
<td>agacataattttatatatactc</td>
</tr>
<tr>
<td>37</td>
<td>aggagaactcatttctcag</td>
<td>38</td>
<td>tcttatagctccgtactga</td>
</tr>
<tr>
<td>39</td>
<td>aatgtatgagtttttatataa</td>
<td>40</td>
<td>tcattctataacaagttatc</td>
</tr>
</tbody>
</table>
FIG. 1: Experimental results after annealing of oligonucleotides. Lane 1 is a molecular size marker for duplexes of 20 to 120 bp in 20 bp increments. Lane 2 contains sequence 5 and its Watson-Crick complement. Lane 3 contains sequences 1-40 after annealing. Lane 4 is sequences 1-40 plus the complement of sequence 5. Lanes 5 and 6 contain sequences 13 and 23, respectively. Lanes 7 and 8 contain oligonucleotides \textit{ggggggggaaaaaacccecccc} and \textit{atgcataagcaagtgc}, respectively, with known secondary structure. There are no detectable duplexes in lane 3, supporting the non-crosshybridizing properties of the designed sequences. By comparing smears in lanes 7 and 8 to well-defined bands in lanes 5 and 6, it is confirmed that designed sequences have little to no secondary structure.

FIG. 2: Lane 1 is a molecular size marker for duplexes of 20 and 40 bp. Lanes 2-41 contain sequences 1-40 by themselves. These results indicate no secondary structure.