Specificity of Hybridization between DNA Sequences Based on Free Energy

Fumiaki Tanaka[1], Atsushi Kameda[2],
Masahito Yamamoto[2,3], and Azuma Ohuchi[2,3]

[1] Graduate School of Engineering, Hokkaido University
North 13, West 8, Kita-ku, Sapporo 060-8628, Japan
[2] CREST, Japan Science and Technology Corporation
4-1-8, Honmachi, Kawaguchi, Saitama, 332-0012, Japan
[3] Graduate School of Information Science and Technology, Hokkaido University
North 14, West 9, Kita-ku, Sapporo 060-0814, Japan
{fumiaki, kameda, masahito, ohuchi}@dna-comp.org

Abstract. We investigated the specificity of hybridization based on a minimum free energy (ΔG_{min}) through gel electrophoresis analysis. The analysis, using 94 pairs of sequences with length 20, showed that sequences that hybridize each other can be separated using the constraint $\Delta G_{min} \leq -14.0$, but cannot be separated using the number of base pairs (BP) in the range from 9 to 18. This demonstrates that the ΔG_{min} is superior to the BP in terms of the capability to separate specific from non-specific sequences. Furthermore, the comparison between sequence design based on ΔG_{min} and that based on the BP, done through a computer simulation, showed that the former outperformed the latter in terms of the number of sequences designed successfully as well as the ratio of successfully designed sequences to the total number of sequences checked.

1 Introduction

Sequence design is an essential step towards success in various applications of DNA computing, including DNA-based computation [1][2] and nano-fabrication [3][4]. Many efforts have been made to design a set of sequences that hybridize only with their complementaries based on the Hamming distance (i.e., the number of base pairs, BP) [5][6][7] or the minimum free energy (ΔG_{min}) [8]. Although many algorithms have been proposed for sequences whose BP or ΔG_{min} values exceed a threshold for satisfactory hybridization specificity, the threshold itself is still unknown. Furthermore, it is not known whether sequence design with the appropriate threshold is best based on the BP or on ΔG_{min}.

We have investigated the specificity of hybridization by analyzing 94 pairs of sequences with length 20 using gel electrophoresis based on the

BP or ΔG_{min}. Based on this experiment, we estimated the thresholds that the BP and ΔG_{min} must reach to enable satisfactory hybridization specificity under regulated conditions such as where two sequences hybridize each other with a 1:1 concentration ratio. We then compared the number of sequences that can be designed based on the BP with that based on ΔG_{min} with these thresholds. Furthermore, through the gel electrophoresis analysis, we found that sequences containing sub-sequence "GGGGG" formed an unintended structure, which appeared to be the four-stranded G4-DNA structure [9]. To confirm that this structure was formed by interaction between GGGGG and GGGGG, we analyzed mutated sequences obtained by changing the sequences with GGGGG.

2 Materials and Methods

Two sequences were hybridized with each other and then analyzed through gel electrophoresis to investigate the hybridization specificity. Because gel electrophoresis does not require an enzyme reaction (e.g., kination and PCR), we can investigate the hybridization specificity while avoiding the influence of extra experimental steps. We checked whether two sequences, A and B, hybridized each other as follows. Sequences A, B, and $A + B$ with a 1:1 concentration ratio were electrophoresed through a 10% polyacrylamide gel. If the band in the lane for $A + B$ corresponded to neither the band in the lane for A nor that for B, we assumed A and B hybridized each other (Figure 1). Thus, if any extra bands in the lane for $A + B$ were observed by eye, we classified the outcome as "hybridize"; otherwise, we classified it as "non-hybridize". However, the double strand between A and B could break down into two single strands during the gel electrophoresis, so we had to take into account that we could not distinguish these from the sequences that did not hybridize with each other. Although this simple model only focuses on the hybridization between two sequences without any competitive sequences, the sequences found to hybridize in the experiment are likely to be harmful even under other conditions. Therefore, it is better to avoid such sequences to avoid blocking a specific hybridization.

The BP between sequences A with length n and B with length m is defined as

$$BP := \min(n, m) - \min_{-m<k<n} H(A, \sigma^k(\overline{B})), \tag{1}$$

where $H(*,*)$ denotes the Hamming distance, σ^k denotes the right (left) bit shift in the case of $k > 0$ ($k < 0$), k denotes the number of the shift, and $\overline{B}$ denotes the reverse complementary of B. Note that the BP is equivalent to the H-measure proposed by Garzon *et al.* [10] in the case of $n = m$.

We calculated ΔG_{min} between two sequences using the extended algorithm for the ΔG_{min} calculation of a single strand [11]. The calculation was done as reported previously [8].

We analyzed 94 pairs of sequences with length 20 having various values of ΔG_{min} for each BP in a range from 9 to 18. The 94 pairs of sequences were chosen as follows. First we randomly generated 100,000 pairs of sequences for each BP through a computer simulation where T_M was in the range $69.58 \leq T_M \leq 72.58$ and the ΔG_{min} between each sequence and itself was greater than or equal to a threshold, -3.0, so that the sequence would not form secondary structures by itself. The T_M values of 69.58 and 72.58 were, respectively, ${T_M}^{ave} - 1.5$ and ${T_M}^{ave} + 1.5$, where ${T_M}^{ave}$ is the average calculated from 10,000 randomly generated sequences with length 20. The frequency distribution curves in Figure 2 show that the number of sequences with a particular BP varies with ΔG_{min}. We then chose pairs of sequences that would contain the maximum and minimum ΔG_{min} value for each BP. When the BP was 12, for example, the selected sequences included those with $\Delta G_{min} = -0.54$ and those with $\Delta G_{min} = -21.24$, respectively the maximum and minimum from 100,000 pairs of sequences.

Oligonucleotides were supplied by Hokkaido System Science and were synthesized using column purification. All oligonucleotides were dissolved in a buffer containing 1 M NaCl, 10 mM Na_2HPO_4, and 1 mM Na_2EDTA with a pH of 7.0. The oligonucleotide concentrations (Ct) of each sample were determined from the difference in absorbance at 260 nm and that at 320 nm using extinction coefficients calculated from dinucleoside monophosphates and nucleotides [12]. The oligonucleotides were hybridized by increasing the temperature to 90 °C for 10 min and lowering the temperature to 20 °C at heating rates of 0.08 and 0.02 °C/s, respectively. It took about 14 and 58 minutes, respectively, to go from 20 °C to 90 °C and from 90 °C to 20 °C: this is almost the typical protocol for the thermodynamic analysis [13]. All gel electrophoresis profiles were obtained using a 10% polyacrylamide gel in a 1×TAE buffer at 200 V for 35 min. We used 2 μl samples at a concentration of 1 μM. Bands in the gels were dyed using SYBR Gold nucleic acid gel stain for 20 min.

3 Experimental Results

3.1 Specificity of Hybridization Based on BP versus ΔG_{min}

Figure 1 shows an example where the BP was 14 with length 20 and the sequences used in this example. A pair of sequences with $\Delta G_{min} = -18.64$ or $\Delta G_{min} = -16.41$ formed double strands resulting in a new band, while that with $\Delta G_{min} = -5.39$ or $\Delta G_{min} = -4.49$ remained two single strands with no extra band. Similar experiments were iterated using 94 pairs of sequences containing the above sequences where the BP was in the range from 9 to 18 with length 20.

The results are shown in Figure 2. All the pairs of sequences that hybridized with each other can be separated from the other pairs by the constraint $\Delta G_{min} \leq -14.0$, but these two groups could not be separated using the BP in the range from 9 to 18. Table 1 shows the number of sequences from 100,000 sequences where $\Delta G_{min} \leq -14.0$ for each BP. The BP had to be less than 13 to guarantee that the number of sequence pairs where $\Delta G_{min} \leq -14.0$ would be lower than 5% of the total. These results demonstrate that ΔG_{min} is superior to the BP in terms of the capability to separate specific from non-specific sequences. However, there seemed to be some pairs of sequences that did not hybridize with each other even though $\Delta G_{min} \leq -14.0$ (e.g., the pair of sequences where $\Delta G_{min} = -15.1$ and the BP was 13). This was probably due to the prediction error for ΔG_{min} and the limit of separability with gel electrophoresis.

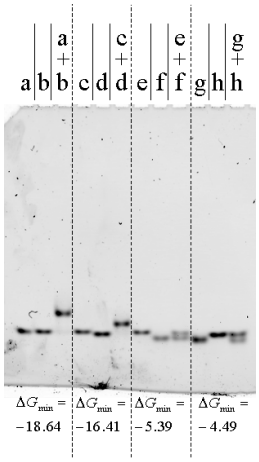
Through the above experiment, we found that five single oligonucleotides resulted in unexpected bands on gels with slow mobility. All of these sequences contained sub-sequence "GGGGG", while the others did not. We believe the sequences containing GGGGG may have formed the four-stranded G4-DNA structures [9].

Table 1. Number of sequences out of 100,000 sequences where $\Delta G_{min} \leq -14.0$ for each BP in the range from 9 to 18.

BP	9	10	11	12	13	14	15	16	17	18
Number of Sequences	46	179	757	2,934	8,544	20,333	41,587	64,716	92,754	99,944

3.2 Sequences Forming Four-Stranded G4-DNA Structures

Sen *et al.* discovered that guanine-rich sequences form four-stranded structures, called G4-DNA, that are linked by Hoogsteen-bonded guanine quar-



Name	Sequence
a	CACAGTCCCGATTTAGCCAG
b	ACTCAACTGGCTAAATCGGG
c	GAGTGCTTGGGGTCAATTTG
d	ATGACCCAAAGCACTCCTTG
e	ACCTCCCCGTTTATTAAGCA
f	TGATTGAGAAAGCGAGAGGT
g	CATTGTGCGGGATTACAAGC
h	GCGTGTAGTGACCCAAAATG

Fig. 1. LEFT: An example of experimental results from the gel electrophoresis. Four sets of sequences, whose BP was 14, were analyzed. The lanes in each set correspond (from left to right) to a sequence A, a sequence B, and sequences $A + B$. RIGHT: Sequences used in the left figure are listed in the direction 5' to 3' from left to right. The letters correspond to lanes from the gel electrophoresis in the left figure.

tets [9]. In particular, they observed that sequences containing GGGGG formed G4-DNA. In addition, the characteristic feature of G4-DNA is its slow electrophoretic mobility, which is consistent with our results. Thus, we think that the unexpected bands, which were observed through the experiment in previous subsection, were due to interaction between GGGGG and GGGGG.

To confirm this, we analyzed five sets of sequences; each set consisted of a sequence with GGGGG, its complementary with CCCCC, and two mutated sequences. The two mutated sequences were generated as follows. One contained GGGG rather than GGGGG, while the other did not contain base G except for GGGGG (Figure 3). For example, in set 'a' in Figure 3, AAGGGGTTCTATGGTGTATT and AGGGGGTTCTATACTCTATT were, respectively, the sequence containing GGGG and the sequence containing no Gs except for GGGGG, where the underlined base is the base changed from the sequence AGGGGGTTCTATGGTGTATT.

Figure 3 shows that sequences with GGGGG formed a structure with slow electrophoretic mobility regardless of the presence of other Gs, while the sequence with GGGG and the sequence with CCCCC did not form such a structure. This indicates that the structures of the unexpected bands were formed by interaction between GGGGG and GGGGG.

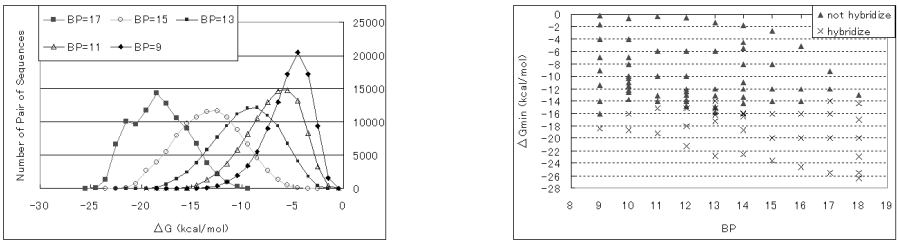


Fig. 2. LEFT: The frequency distribution curve of 100,000 sequences with length 20 for each odd-numbered BP from 9 to 18. RIGHT: Specificity of hybridization based on BP versus ΔG_{min} from the gel electrophoresis analysis using 94 pairs of sequences.

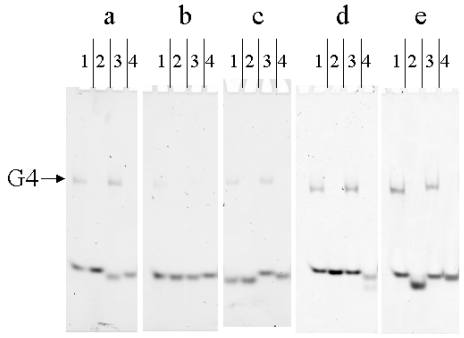
The structures of the unexpected bands, which we believe are G4-DNA, must compete with the specific hybridization and will be intermediate to the unintended structures. Therefore, we conclude that sequences with GGGGG should be avoided when designing specific sequences.

3.3 Sequence Design Based on ΔG_{min} versus That Based on BP

To evaluate sequence design based on ΔG_{min}, we compared it with sequence design based on the BP in terms of the number of sequences successfully designed within 10 hours. We designed sequences with length 20 such that $69.58 \leq T_M \leq 72.58$ and $\Delta G_{min} > \Delta G^*_{min}$ ($BP < BP^*$) in the combinations described below, where ΔG^*_{min} and BP^* are the thresholds. We set $\Delta G^*_{min} = -14.0$ and BP^* =11, 12, or 13 based on Figure 2 and Table 1; for BP =11, 12, or 13, there were, respectively, 757, 2,934, or 8,544 pairs of sequences (out of the 100,000 pairs of sequences) where $\Delta G_{min} \leq -14.0$. In the case that n sequences were to be designed, the combinations to be considered for the ΔG_{min} (BP) calculation were as follows.

1. $< U_i, U_j U_k >$ $(0 \leq i, j, k \leq n-1)$
2. $< U_i, U_j V_k >$ $(0 \leq i, j, k \leq n-1)$, $i \neq k$
3. $< U_i, V_j U_k >$ $(0 \leq i, j, k \leq n-1)$, $i \neq j$
4. $< U_i, V_j V_k >$ $(0 \leq i, j, k \leq n-1)$, $(i \neq j) \wedge (i \neq k)$,

where U_i is the $i-th$ sequence, V_j is the complementary of U_j, $X_j X_k$ ($X_j \in \{U_j, V_j\}$, $X_k \in \{U_k, V_k\}$) is the concatenation of sequences X_j and X_k in that order, and $< U_i, X_j X_k >$ is the combination of sequences U_i and $X_j X_k$. For example, if $U_i = CCCCC$, $U_j = AGAGA$, and $U_k = TCTCT$, $< U_i, U_j U_k >$ means the combination of sequences



Set	No.	Sequence
a	1	A**GGGGG**TTCTATGGTGTATT
	2	A<u>A</u>GGGGTTCTATGGTGTATT
	3	A**GGGGG**TTCTAT<u>AC</u>T<u>C</u>TATT
	4	AATACACCATAGAACCCCCT
b	1	AAAGTTCTCAAAAGA**GGGGG**
	2	AAAGTTCTCAAAAGAGGGG<u>C</u>
	3	AAA<u>C</u>TTCTCAAAA<u>C</u>A**GGGGG**
	4	CCCCCTCTTTTGAGAACTTT
c	1	TCTTGTTATCTCGTA**GGGGG**
	2	TCTTGTTATCTCGTA<u>A</u>GGGG
	3	TCTT<u>A</u>TTATCTC<u>A</u>TA**GGGGG**
	4	CCCCCTACGAGATAACAAGA
d	1	CTCTTGT**GGGGG**TGTATTTT
	2	CTCTTGT<u>A</u>GGGGTGTATTTT
	3	CTCTT<u>A</u>T**GGGGG**T<u>A</u>TATTTT
	4	AAAATACACCCCCACAAGAG
e	1	CCTTAACATTCTA**GGGGGG**T
	2	CCTTAACATTCTAGGGG<u>AA</u>T
	3	CCTTAACATTCTA<u>C</u>**GGGGG**T
	4	ACCCCCCTAGAATGTTAAGG

Fig. 3. LEFT: Five sets of sequences consisting of a sequence with GGGGG and mutated sequences were analyzed. In each set, 1: sequence with GGGGG; 2: sequence with GGGG; 3: sequence without G except for GGGGG; 4: complementary of sequence 1 with CCCCC. RIGHT: Sequences used in the left figure are listed in the direction 5' to 3' from left to right. The letters and numbers correspond to lanes for the gel electrophoresis in the left figure. Sequence GGGGG is shown in boldface. The underlined bases show mutated bases from the sequence with GGGGG.

CCCCC and *AGAGATCTCT*. The algorithm for both sequence designs was a random generate-and-test algorithm that generated a candidate sequence randomly and tested whether the sequence satisfied the constraints. Furthermore, when we designed sequences based on ΔG_{min}, we used the ΔG_{gre} filtering method, which effectively excluded inappropriate sequences where $\Delta G_{min} \leq \Delta G^*_{min}$, thereby reducing the computation time. Finally, the sequence design based on ΔG_{min} checked the candidate sequence with the T_M, ΔG_{gre}, and ΔG_{min} filters in that order, while that based on BP checked the candidate sequence with T_M and BP filters in that order (see reference [8] for details). The computational experiments were performed using Turbolinux Workstation 7.0 on a computer with a Pentium 4 2.26-GHz CPU and 256 MB of memory. The experiments were iterated five times with a different seed for the random generator. The results are shown in Table 2. The number of sequences successfully

designed based on ΔG_{min} exceeded that based on the BP. This shows that sequence design based on ΔG_{min} is more effective than that based on the BP when designing specific sequences that hybridize with only the complementary.

Table 2. Number of sequences successfully designed within 10 hours based on ΔG_{min} versus BP. The experiments were iterated five times with a different seed for the random generator. In the column $\Delta G_{min} > -14.0$, the numbers in parentheses correspond to the design strategy without ΔG_{gre} filtering. Using ΔG_{gre} filtering enables the design of more sequences. Sequence design based on ΔG_{min} outperformed that based on the BP even without ΔG_{gre} filtering.

Trial	$\Delta G_{min} > -14.0$	$BP < 11$	$BP < 12$	$BP < 13$
1	106 (92)	11	27	64
2	106 (90)	11	27	65
3	96 (87)	13	27	60
4	103 (87)	12	24	62
5	104 (87)	10	25	62
Average	103 (88.6)	11.4	26	62.6
Standard Deviation	4.1 (2.3)	1.1	1.4	1.9

3.4 Comparison between the Solution Space Based on ΔG_{min} and That Based on BP

Above we demonstrated that more sequences can be successfully designed based on ΔG_{min} than based on the BP. However, the number of sequences that can be designed also depends on the sequence design algorithm. Thus, one might think that sequence design based on the BP with a more sophisticated algorithm might outperform that based on ΔG_{min}. It is difficult to prove that any and all algorithms based on ΔG_{min} are superior to those based on the BP. Instead, we investigated the ratio of successfully designed sequences to the total number of sequences checked because this ratio corresponds to the size of the solution space that can be designed under predefined constraints. Table 3 shows that the ratio of sequences successfully designed based on ΔG_{min} was far larger than that based on the BP (e.g., $2.8\% \gg 1.9 \cdot 10^{-4}\%$). This means that the solution space that can be designed based on ΔG_{min} is undoubtedly larger than that based on the BP. Therefore, although the time complexity of BP is less than that of ΔG_{min}, the number of sequences that can be designed based on ΔG_{min} is greater than that for the BP (Table 2). These results demonstrate the rationality of sequence design based on ΔG_{min}.

Table 3. Ratio of successfully designed sequences to total number of sequences checked. The experiments were iterated five times with a different seed for the random generator. In the column $\Delta G_{min} > -14.0$, the numbers in parentheses correspond to the design strategy without ΔG_{gre} filtering.

Trial	$\Delta G_{min} > -14.0$	$BP < 11$	$BP < 12$	$BP < 13$
1	3.0 (5.1) %	$4.0 \cdot 10^{-6}$ %	$2.3 \cdot 10^{-5}$ %	$2.0 \cdot 10^{-4}$ %
2	2.8 (4.9) %	$1.6 \cdot 10^{-6}$ %	$1.2 \cdot 10^{-5}$ %	$2.2 \cdot 10^{-4}$ %
3	2.4 (4.5) %	$5.9 \cdot 10^{-6}$ %	$1.4 \cdot 10^{-5}$ %	$1.9 \cdot 10^{-4}$ %
4	3.2 (4.2) %	$1.8 \cdot 10^{-6}$ %	$2.3 \cdot 10^{-5}$ %	$1.7 \cdot 10^{-4}$ %
5	2.6 (5.1) %	$1.7 \cdot 10^{-6}$ %	$1.1 \cdot 10^{-5}$ %	$1.9 \cdot 10^{-4}$ %
Average	2.8 (4.8) %	$3.0 \cdot 10^{-6}$ %	$1.7 \cdot 10^{-5}$ %	$1.9 \cdot 10^{-4}$ %
Standard Deviation	0.3 (0.4) %	$1.9 \cdot 10^{-6}$ %	$0.6 \cdot 10^{-5}$ %	$0.2 \cdot 10^{-4}$ %

4 Conclusions

We conclude that using ΔG_{min} is preferable to using the BP to separate specific hybridization from non-specific hybridization. With an appropriate threshold, sequence design using ΔG_{min} outperformed that using the BP in terms of the number of sequences that could be successfully designed. Comparison of the ratio of successfully designed sequences to the total number of sequences checked showed that the superiority of ΔG_{min} over the BP probably does not depend on the algorithm used for the sequence design.

In addition, our analysis of sequences with GGGGG and their mutated sequences suggested that the sequences with GGGGG formed G4-DNA. Thus, sequences with GGGGG should be avoided when designing specific sequences.

References

1. D Faulhammer, AR Cukras, RJ Lipton, and LF Landweber. Molecular computation: RNA solutions to chess problems. *Proc Natl Acad Sci U S A*, 97(4):1385–9, Feb 2000.
2. Ravinderjit S Braich, Nickolas Chelyapov, Cliff Johnson, Paul W K Rothemund, and Leonard Adleman. Solution of a 20-variable 3-SAT problem on a DNA computer. *Science*, 296(5567):499–502, Apr 2002.
3. Hao Yan, Xiaoping Zhang, Zhiyong Shen, and Nadrian C Seeman. A robust DNA mechanical device controlled by hybridization topology. *Nature*, 415(6867):62–5, Jan 2002.
4. William M Shih, Joel D Quispe, and Gerald F Joyce. A 1.7-kilobase single-stranded DNA that folds into a nanoscale octahedron. *Nature*, 427(6975):618–21, Feb 2004.
5. M. Arita and S. Kobayashi. DNA sequence design using templates. *New Generation Computing*, 20:263–277, 2002.

6. D. Tulpan, H. Hoos, and A. Condon. Stochastic local search algorithms for DNA word design. *Proceeding of 8th International Workshop on DNA-Based Computers, LNCS*, 2568:229–241, 2002.
7. Satoshi Kashiwamura, Atsushi Kameda, Masahito Yamamoto, and Azuma Ohuchi. Two-step search for DNA sequence design. *IEICE TRANSACTIONS on Fundamentals of Electronics, Communications and Computer Sciences Special Section on Papers Slected from 2003 International Technical Conference on Circuts/Systems, Computer and Communications (ITC-CSCC 2003)*, E87-A(6):1446–1453, 2004.
8. Fumiaki Tanaka, Atsushi Kameda, Masahito Yamamoto, and Azuma Ohuchi. Design of nucleic acid sequences for DNA computing based on a thermodynamic approach. *Nucleic Acids Res*, 33(3):903–11, 2005.
9. D Sen and W Gilbert. Formation of parallel four-stranded complexes by guanine-rich motifs in DNA and its implications for meiosis. *Nature*, 334(6180):364–6, Jul 1988.
10. M. Garzon, R. Deaton, P. Neather, D. R. Franceschetti, and R. C. Murphy. A new metric for DNA computing. In *Poceedings of 2nd Annual Genetic Programming Conference*, volume GP-97, pages 472–8, 1997.
11. M Zuker and P Stiegler. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res*, 9(1):133–48, Jan 1981.
12. DM Gray, SH Hung, and KH Johnson. Absorption and circular dichroism spectroscopy of nucleic acid duplexes and triplexes. *Methods Enzymol*, 246:19–34, 1995.
13. Fumiaki Tanaka, Atsushi Kameda, Masahito Yamamoto, and Azuma Ohuchi. Thermodynamic parameters based on a nearest-neighbor model for DNA sequences with a single-bulge loop. *Biochemistry*, 43(22):7143–50, Jun 2004.