Design of Code Words for DNA Computers and Nanostructures with Consideration of Hybridization Kinetics

Tetsuro Kitajima, Masahiro Takinoue, Ko-ichiroh Shohda, and Akira Suyama

Department of Life Science and Institute of Physics,
Graduate school of Arts and Sciences
The University of Tokyo
3-8-1 Komaba, Meguro-ku, Tokyo 153-8902, Japan
{kitajima, takinoue}@genta.c.u-tokyo.ac.jp, suyama@dna.c.u-tokyo.ac.jp

Abstract. We have developed a method for designing rapidly-hybridizing orthonormal DNA sequences. Two conditions were used in the prediction method. One condition concerned the stability of the self-folded secondary structures of forward and reverse strands. The other condition concerned the nucleation capability of complementary strands at the tails of their self-folded secondary structures. These conditions were derived from the complementary strands' experimentally-determined hybridization rates' dependence on their stability and nucleation capability. These dependences were examined for 37 orthonormal DNA sequences randomly selected from our set of 300 orthonormal DNA sequences. By applying this new method to the set of 300 orthonormal DNA sequences, more than 100 rapidly-hybridizing sequences were obtained.

1 Introduction

DNA computing and DNA nanotechnology employ remarkable features unique to DNA and RNA molecules. The interactions and structures of DNA and RNA molecules can be most successfully designed in terms of their base sequences among various molecules. In DNA computing, programs and data are encoded into DNA/RNA sequences, while in DNA nanotechnology, structures and functions are encoded into DNA/RNA sequences. Thus, the design of DNA and RNA sequences is a crucial step for DNA computing and DNA nanotechnology.

Various methods have been developed to design DNA and RNA sequences for DNA computing and DNA nanotechnology[1,2,3,4,5,6,7,8,9]. Sets of DNA sequences of a uniform length and stability without mis-hybridizations and stable self-folded structures have been designed and applied to DNA computing, DNA probe sensors for genome analysis, and constructions of DNA nanostructures and nanodevices. The design methods so far developed are based on thermodynamic models. The stability of desirable and undesirable hybrids formed through intermolecular base-pairing, and the stability of self-folded structures formed through

M.H. Garzon and H. Yan (Eds.): DNA 13, LNCS 4848, pp. 119–129, 2008.

intramolecular base-pairing are calculated from base sequences by using thermodynamic models and parameters.

In DNA computing and DNA nanotechnology, however, not only thermodynamic properties but also the kinetic properties of DNA/RNA sequences substantially affect the results of computations and nanostructure constructions. Non-uniform hybridization rates will make the speed at which instruction codes are executed dependent on the content of the data, especially for computation on autonomous DNA computers running under isothermal conditions. This data-dependent execution speed makes the computation less reliable, while the non-uniformity of the rates will also make the construction of DNA nanostructures more complicated. Therefore, design methods in which kinetic properties are also considered are essential for the further development of DNA computing and DNA nanotechnology.

In this study, we have explored the DNA sequence dependence of hybridization rates in order to develop a DNA sequence design method that takes the kinetic properties of DNA/RNA hybridization into consideration. The rate of hybridization of two complementary strands was measured for 37 DNA sequences randomly chosen from our set of orthonormal DNA sequences 23 nucleotides long. The orthonormal sequences were designed to be orthogonal and normalized. This orthogonality means that every sequence in the set significantly hybridizes neither with any sequences in the set other than its compliment nor with any concatenated sequences made of sequences in the set without its complement. Normalization in this context means that every sequence has a uniform length and a uniform duplex stability, and has no very stable self-folded secondary structures that may significantly hinder rapid hybridization with its complement. The orthonormal DNA sequences were assured to have a uniform duplex stability not only by an equal melting temperature but also by the equal free energy changes accompanying duplex formation. However, the hybridization rate significantly depended on the DNA sequences. The stability and the shape of the sequences' self-folded secondary structures were examined to elucidate their relationships to the hybridization rates. Based on these relationships, a method to design a DNA sequence for rapidly-hybridizing orthonormal DNA sequences has been proposed.

2 Materials and Methods

2.1 DNA Sequences

The DNA sequences used in hybridization experiments and secondary structure predictions were 37 orthonormal DNA sequences randomly selected from our 23-mer orthonormal sequence set containing more than 300 DNA sequences. Their sequences were as follows: 5'-GCATCTACACTCAATACCCAGCC-3' , 5'-CGTCTATTGCTTGTCACTTCCCC-3' , 5'-GGCTCTATACGATTAAACTCCCC-3' , 5'-GAAGGAATGTTAAAATCGTCGCG-3' , 5'-GCACCTCCAAATAAAAACTCCGC-3' , 5'-GAGAAGTGCTTGATAACGTGTCT-3' , 5'-GCATGTGTAGTTATCAGCTTCCA-3' , 5'-CTAGTCCATTGTAACGAAGGCCA-3' , 5'-GTCCCGGAAAATACTATGAGACC-3' ,

5'-GAGTCCGCAAAAATATAGGAGGC-3' , 5'-CATCTGAACGAGTAAGGACCCCA-3' ,
5'-CGCGATTCCTATTGATTGATCCC-3' , 5'-GGTGGCTTATTTACAGGCGTTAG-3' ,
5'-TTCGGTTCTCTCCAAAAAAAGCA-3' , 5'-GGCGCTTAAATCATCTTTCATCG-3' ,
5'-CCGTCGTGTTATTAAAGACCCCT-3' , 5'-CGAGAGTCTGTAATAGCCGATGC-3' ,
5'-TGGCACTTATAGCTGTCGGAAGA-3' , 5'-GGCTGTTTACAAAATCGAGCTAG-3' ,
5'-TGCGAAATTTGAAAAATGGCTGC-3' , 5'-GCATTGAGGTATTGTTGCTCCCA-3' ,
5'-GGCTGTCAATTTATCAGGGAGGC-3' , 5'-GCCTCAAGTACGACTGATGATCG-3' ,
5'-GAAGCCCTATTTTGCAATTCCCC-3' , 5'-CGCGGGTACGTTGATGTAACAAA-3' ,
5'-ATGGGAACCTAAAAGTGTGGCTA-3' , 5'-GAGTCAATCGAGTTTACGTGGCG-3' ,
5'-TTCGCTGATTGTAGTGTTGCACA-3' , 5'-GCCTCACATAACTGGAGAAACCT-3' ,
5'-CCATCAGGAATGACACACACAAA-3' , 5'-GGGATAGAACTCACGTACTCCCC-3' ,
5'-CCATATCCGATTATTAGCGACGG-3' , 5'-GGGATCAGTTGTACACTCCCTAG-3' ,
5'-CTGTGATGATACCGTTCTTCACC-3' , 5'-CGCGGTTGAAATAACTAATCGCG-3' ,
5'-GGTCGAAACGTTATATTAACGCG-3' , 5'-TAGCACCCGTTAAAACGGAAATG-3' .

The DNA strands of 38 orthonormal sequences and their complements were synthesized and purified by HPLC commercially (SYGMA genosys, Japan).

2.2 DNA Hybridization

The time course of the hybridization of two complementary DNA strands was measured on a fluorescence spectrophotometer LS 55 (Perkin Elmer, USA) equipped with a stopped-flow apparatus RX-2000 (Applied Photophysics, UK). Two complementary DNA strands at 50 nM each in 1xSSC (0.015 M $Na_3 -$ citrate, 0.15 M NaCl) containing PicoGreen® (Invitrogen, USA) were mixed rapidly (a dead-time of 8 ms) at 25°C through the use of the stopped-flow apparatus. The DNA duplex formation was followed by the fluorescence emission from PicoGreen® at 523 nm exited at 502 nm. PicoGreen® is a dye for quantitating double-stranded DNA (dsDNA) in the presence of single-stranded DNA/RNA. The linear detection range of the dye extends over more than four orders of magnitude in dsDNA concentration (from 25 pg/ml to 1,000 ng/ml), allowing the precise observation of the time course of DNA hybridization.

2.3 Determination of Hybridization Rates

The time courses of the hybridization of two complementary DNA strands were fitted to a single-exponential model:

$$I(t) = I_\infty + A \exp(-t/\tau),$$

where $I(t)$ is the observed fluorescence intensity at time t , I_∞ is the final fluorescence intensity, A is the amplitude of hybridization, and $1/\tau$ is the rate of hybridization. The best-fit-values of the hybridization rates were obtained by nonlinear regression using Origin 7.0 and Microsoft Excel solver. The goodness of fit was measured by the value of R2, and was also confirmed visually through every plot overlapping the observed data and the best-fit-curve.

2.4 Secondary Structure Prediction

The self-folded secondary structures of the single DNA strands of 37 orthonormal sequences were predicted by using the mfold system, which is a tool for predicting RNA/DNA secondary structures by free energy minimization. For the prediction, a condition of 0.195 M Na^+ and 0 M Mg^{2+} was employed. This condition is equivalent to the 1xSSC at which the hybridization experiments were performed.

3 Results and Discussion

3.1 Hybridization Rates of Orthonormal DNA Sequences

The hybridization rate of two complementary DNA strands was determined for 37 orthonormal DNA sequences randomly selected from our set of 300 orthonormal sequences 23 nucleotides long. The time course of the hybridization was followed by measurement of the fluorescence intensity, which is proportional to the concentration of hybrid duplexes formed (i.e., to that of the base-pair stacks formed). The observed fluorescence intensity data were fitted to a single-exponential model to determine the hybridization time τ. Figure 1 shows a typical time course of the hybridization and the best-fitting single exponential curve. For all of the 37 orthonormal sequences, the hybridization rate was determined from three independent hybridization experiments, which were highly reproducible.

Figure 2 shows the distribution of the hybridization time determined for the 37 orthonormal sequences. The rate of hybridization significantly depended on DNA sequences although all DNA duplexes have a uniform thermodynamic stability. Most of the sequences finished hybridizing in 200 s, but some of the sequences indicated a very slow hybridization (i.e., a hybridization time of more than 15 min).

3.2 Hybridization Rates and the Stability of Self-folded Secondary Structures

The hybridization rates of complementary strands of nucleic acids are affected by the stability of the strands' self-folded structures. The strands' stable secondary structures are shown to significantly reduce the rate of hybridization [10,11]. Therefore, those sequences that form very stable secondary structures were discarded in the design of the set of 300 orthonormal DNA sequences. However, the threshold value of the free energy change of secondary structure formation below which sequences should be discarded was not evident, so that the set contains secondary structure sequences with a wide range of stabilities: for some sequences the value of the free energy change in secondary structure formation ΔG is positive, and for some other sequences ΔG is as low as -5 kcal/mol. Consequently, the slow hybridization observed for some sequences may be due to the formation of slightly stable secondary structures. Thus we examined the dependence of the observed hybridization time τ on the stability of predicted secondary structures.

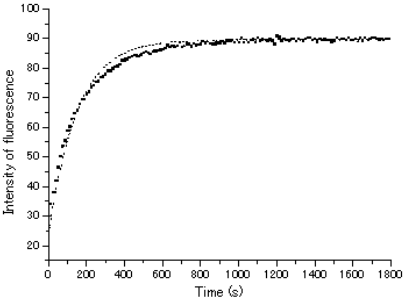

Fig. 1. Typical time course of the hybridization of orthonormal DNA sequence. Two complementary DNA strands at 50 nM each were rapidly mixed at 25 °C in 1xSSC in the presence of PicoGreen®. The sequence of DNA was 5f-CCATATCCgATTATTAgCgACgg-3f. The closed squares are observed fluorescence intensities and the broken line is the best-fitting single-exponential curve obtained by non-linear regression. The hybridization time, which is the reciprocal of the hybridization rate, determined by the best-fitting curve was 1.5×10^2 s.

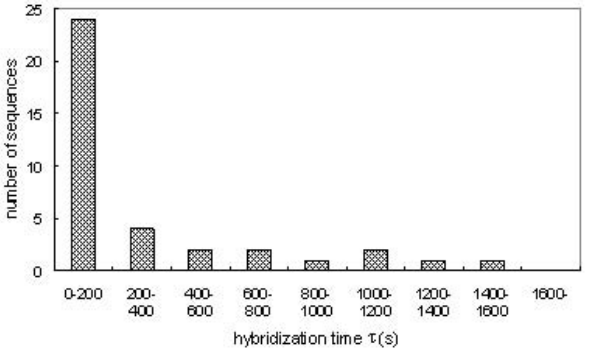

Fig. 2. Distribution of hybridization times of 37 orthonormal DNA sequences

Figure 3 shows the relationship between the stability of self-folded secondary structure predicted using the mfold method and the observed hybridization time τ. The stability of the secondary structure was measured in terms of the free energy change ΔG in secondary structure formation. When the value of ΔG is positive, the secondary structure is less stable than the unstructured coil. When the value of ΔG is negative, the secondary structure is more stable than the unstructured coil. Only the lowest energy structures were considered in this study, though some sequences were predicted to have more than one possible structure. Figure 3 indicates that two complementary strands, i.e., a forward and a reverse strand, hybridized rapidly ($\tau < 240$ s) when either of the two strands had a positive value of ΔG. Especially when both of them had a positive ΔG value, the hybridization was more rapid ($\tau < 120$ s). When either of the two strands

had a negative ΔG value, some of the DNA sequences showed slow hybridization rates, while the other sequences still hybridized rapidly. Therefore, an increase in the stability of the self-folded secondary structures actually decreased the hybridization rates of two complementary strands. However, there must exist factors that affect the rate of hybridization other than the stability of secondary structures, because some DNA sequences with largely negative ΔG values still hybridized rapidly.

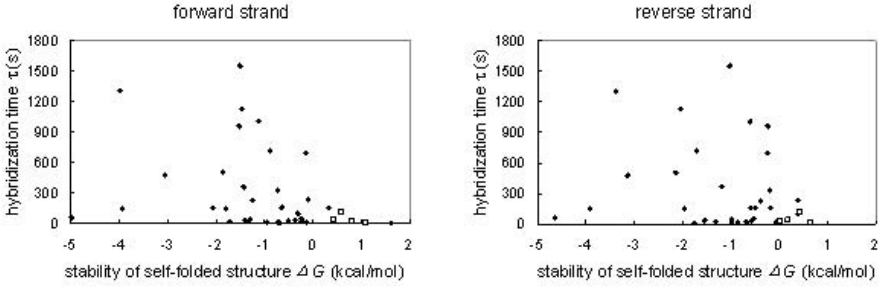

Fig. 3. Dependence of the hybridization time τ on the stability of self-folded structures of forward and reverse strands. The stability of the structure was measured in terms of the predicted free energy change ΔG in secondary structure formation. Open squares designate data of the orthonormal DNA sequences that have positive ΔG values for both forward and reverse strands.

3.3 Hybridization Rates and the Nucleation Capability of Self-folded Structures

The hybridization of nucleic acid strands requires the formation of a nucleus composed of at least three contiguous base-pairs. As soon as nucleation occurs, each duplex zips up to completion instantly. In the hybridization of short DNA strands at low concentrations such as those studied here, it has long been accepted that the nucleation step is rate-limiting [12]. We thus examined how the nucleation step affects the hybridization rate of complementary strands of the 37 orthonormal DNA sequences.

In the process of nucleation, each strand tries to find unpaired-base stretches of complementary sequences. Those stretches are found only in the loop and tail (end-coil) regions of self-folded strands. The length of the orthonormal sequences studied here was as short as 23 nucleotides, so that the size of the loops found in their secondary structures may not be large and flexible enough to perform a rapid nucleation. Tails are, in contrast, more flexible, so that unpaired bases found in the tails would more easily be involved in nucleation. We thus focused on the length of tails and examined whether the self-folded secondary structures of the orthonormal sequences have tails that are long enough for nucleation.

The orthonormal DNA sequences were classified into two groups, 'nucleation-inhibited' and 'nucleation-allowed' sequences, according to the length of contiguous unpaired-bases at the ends. A DNA sequence was defined as 'nucleation-

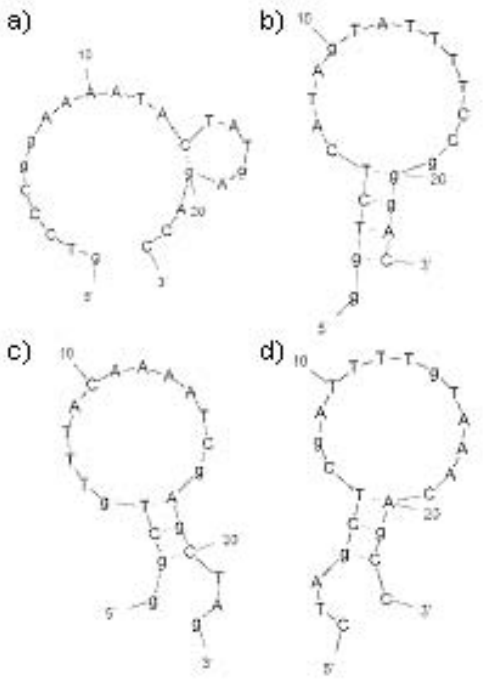

Fig. 4. Self-folded structures of forward and reverse strands of two orthonormal DNA sequences A and B. The most stable secondary structures of the forward (a) and the reverse (b) strand of sequence A. Those of the forward (c) and the reverse (d) strand of sequence B.

inhibited' if neither tail at the 5f- nor the 3f-end of its forward strand can be involved in the formation of a nucleus of 3 or 4 base-pairs. A DNA sequence was defined as 'nucleation-allowed' if either tail at the 5f- or the 3f-end of its forward strand can be involved in nucleation. The orthonormal DNA sequence shown in Figures 4a and 4b, for example, is nucleation-inhibited. The 5f- end of the forward strand has 13 contiguous unpaired-bases, which is long enough to form a nucleus of 3 or 4 base-pairs (Fig. 4a). However, the 5f-end cannot be involved in nucleation because the 3f-end of the reverse strand has no unpaired -base (Fig. 4b). The 3f-end of the forward strand cannot be involved in nucleation either because the 3f-end of the forward strand has 3 contiguous unpaired-bases and the 5f-end of the reverse strand has only one unpaired-base. Therefore, for this orthonormal DNA sequence nucleation is inhibited. On the other hand, the orthonormal DNA sequence shown in Figures 4c and 4d is nucleation-allowed. The 5f- end of the forward strand has one unpaired base and the 3f-end of the reverse strand has also one unpaired base (Figs. 4c and 4d). The 5f-end of the forward strand, therefore, cannot be involved in nucleation. However, the 3f-end can be involved in the

formation of a nucleus of 3 base-pairs because the 3f-end of the forward strand and the 5f-end of the reverse strand, both ends have 3 contiguous unpaired-bases. Therefore, for this orthonormal sequence nucleation is allowed.

Figure 5 indicates how the number of nucleation-inhibited and nucleation-allowed sequences varied with the hybridization time τ. The number of nucleation-inhibited sequences increased with the increase in hybridization time. In contrast, the number of the nucleation-allowed sequences increased as the hybridization time decreased. Therefore, nucleation at the tails of self-folded secondary structures should be one of the critical factors affecting the hybridization rate.

3.4 Prediction of Orthonormal DNA Sequences Rapidly Hybridizing with Complementary Strands

The effect of the self-folded secondary structures' thermodynamic stability on the hybridization rate (Fig. 3) and that of the nucleation capability at the tails of self-folded secondary structures on the hybridization rate (Fig. 5) have provided the basic concept of a method for the design of DNA code word sequences rapidly hybridizing with complementary strands. Each of these effects by itself was not sufficiently significant to determine the hybridization rate. Therefore, in the design method both of these factors were taken into consideration.

The table summarizes how the number of rapidly-hybridizing orthonormal DNA sequences and the number of slowly-hybridizing ones depended on the self-folded secondary structures' thermodynamic stability and the nucleation capability at the tails of the self-folded secondary structures. In the table, DNA sequences with a hybridization time of less than 300 s are categorized as rapidly-hybridizing sequence and those with 300 s or more are categorized as slowly-hybridizing sequences. The hybridization-time threshold of 300 s was determined according to the experimental conditions of our autonomous DNA computing system RTRACS, and thus it is not exclusive to this design method.

From the results shown in the table, two conditions were derived to predict rapidly-hybridizing orthonormal DNA sequences. In Condition 1, both the forward and reverse strands of a sequence have positive ΔG values. Condition 1 self-evidently assures that the sequence is nucleation-allowed. In Condition 2, the sum of the ΔG values of a sequences' forward and reverse strands is larger than -1 kcal/mol and the sequence is also nucleation-allowed. Condition 2 is always applied after Condition 1; that is, Condition 2 is applied only to those DNA sequences whose forward or reverse strands have a negative ΔG value.

By applying Condition 1 to the set of 37 orthonormal DNA sequences studied here, 4 sequences were selected. Then by applying Condition 2, 12 sequences were further selected. A total of 16 sequences out of the 37 orthonormal DNA sequences were predicted as rapidly-hybridizing sequences. According to the table, only one sequence out of the 16 sequences predicted is not a rapidly-hybridizing sequence, while its hybridization time (330 s) was close to the threshold time (300 s). The false positive rate of prediction, therefore, is as small as 6%. Our orthonormal DNA sequence set contains 300 sequences. We then applied

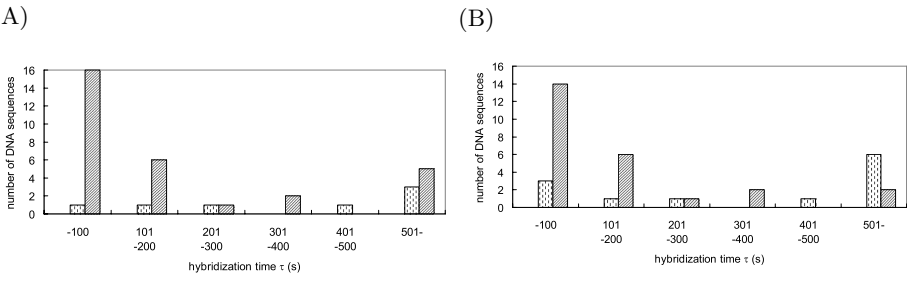

Fig. 5. Change of the number of the nucleation-inhibited and the nucleation-allowed orthonormal DNA sequences with the hybridization time τ . Hatched bars indicate the number of the nucleation-allowed sequences, and the dashed bars that of nucleation-inhibited sequences. The size of the nucleus allowed to form is 3 base-pairs (a) and 4 base-pairs (b).

Table 1. Summary of the dependence of the number of orthonormal DNA sequences on the stability of self-folded secondary structures and the nucleation of complementary strands at their tails

$\Delta G_{fwd} + \Delta G_{rev}$	nucleation-inhibited		nucleation-allowed	
	hybridization rate		hybridization rate	
(kcal / mol) (#)	slow	rapid	slow	rapid
Case 1: Both of ΔG_{fwd} and ΔG_{rev} are positive				
	0	0	**0**	**4**
Case 2: Either of ΔG_{fwd} or ΔG_{rev} is negative				
$+2 \sim +1$	0	0	**0**	**1**
$+1 \sim 0$	0	0	**0**	**2**
$0 \sim -1$	1	0	**1**	**8**
$-1 \sim -2$	0	1	2	1
$-2 \sim -3$	0	1	3	4
$-3 \sim -4$	1	0	0	2
$-4 \sim -5$	0	0	1	0
$-5 \sim -6$	0	0	0	0
$-6 \sim -7$	1	0	0	0
$-7 \sim -8$	1	0	0	1
$-8 \sim -9$	0	0	0	0
$-9 \sim -10$	0	0	0	1

#) ΔG_{fwd} and ΔG_{rev} stand for ΔG of a forward strand and that of a reverse strand, respectively.

both Conditions 1 and 2 to the set of 300 orthonormal DNA sequences. A set of 108 rapidly-hybridizing sequences, which may contain 7-8 slowly-hybridizing sequences, was obtained.

The present prediction method using Conditions 1 and 2 may be satisfactory because 108 sequences are sufficient for most studies using rapidly-hybridizing orthonormal DNA sequences. In a set of 300 orthonormal DNA sequences, however, many sequences may still be predicted as slowly-hybridizing while they could actually be hybridizing rapidly, because the table contains many nucleation-allowed sequences hybridizing rapidly with largely negative ΔG values of less than -1 kcal/mol. If those sequences can be distinguished from other nucleation-allowed sequences hybridizing slowly by using additional conditions, the predictability of the method will be much increased. One promising condition would concern the stability of short duplexes adjacent to the tails involved in nucleation. It is conceivable that even when a sequence has a largely negative ΔG value indicating a globally-stable self-folded secondary structure, its complementary strands should hybridize rapidly if the adjacent short duplexes are unstable. Such an additional condition would increase the number of rapidly-hybridizing sequences predicted by the method as keeping the false positive rate substantially low.

4 Conclusion

We have developed a method for designing rapidly-hybridizing orthonormal DNA sequences. Two conditions are used in the prediction method. One condition concerns the stability of the self-folded secondary structures of forward and reverse strands, while the other concerns the nucleation at the tails of their self-folded secondary structures. More than 100 rapidly-hybridizing orthonormal DNA sequences were obtained by the present prediction method.

Acknowledgements

This work was supported by a grant for SENTAN (Development of System and Technology for Advanced Measurement and Analysis) from the Japan Science and Technology Agency (JST), and by a grant-in-aid for the 21st Century COE program “Research Center for Integrated Science” and for Scientific Research on Priority Areas “Molecular Programming” from the Ministry of Education, Culture, Sports, Science, and Technology of Japan.

References

1. Tulpan, D., Andronescu, M., Chang, S.B., Shortreed, M.R., Condon, A., Hoos, H.H., Smith, L.M.: Thermodynamically based DNA strand design. Nucleic Acids Res. 33, 4951–4964 (2005)
2. Shortreed, M.R., Chang, S.B., Hong, D., Phillips, M., Campion, B., Tulpan, D.C., Andronescu, M., Condon, A., Hoos, H.H., Smith, L.M.: A thermodynamic approach to designing structure-free combinatorial DNA word sets. Nucleic Acids Res. 33, 4965–4977 (2005)
3. Dirks, R.M., Lin, M., Winfree, E., Pierce, N.A.: Paradigms for computational nucleic acid design. Nucleic Acids Res. 32, 1392–1403 (2004)

4. Arita, M., Kobayashi, S.: Sequence design using template. New Generation Computing 20, 263–277 (2002)
5. Jonoska, N., Kephart, D., Mahalingam, K.: Generating DNA code words. Congressus Numerantium 156, 99–110 (2002)
6. Penchovsky, R., Ackermann, J.: DNA library design for molecular computation. J. Comp. Biol. 10, 215–229 (2003)
7. Feldkamp, U., Rauhe, H., Banzhaf, W.: Software Tools for DNA Sequence Design. Genetic Programming and Evolvable Machines 4, 153–171 (2003)
8. Garzon, M., Deatorthonormal, J.: Codeword design and information encoding in DNA ensembles. Natural Computing 3, 253–292 (2004)
9. Kari, L., Konstantinidis, S., Sosík, P.: Preventing undesirable bonds between DNA codewords. Lect. Notes Comput. Sc. 3384, 182–191 (2005)
10. Kushon, S.A., Jordan, J.P., Seifert, J.L., Nielsen, H., Nielsen, P.E., Armitage, B.A.: Effect of secondary structure on the thermodynamics and kinetics of PNA hybridization to DNA hairpins. J. Am. Chem. Soc. 123, 10805–10813 (2001)
11. Gao, Y., Wolf, L.K., Georgiadis, R.M.: Secondary structure effects on DNA hybridization kinetics: a solution versus surface comparison. Nucleic Acids Res. 34, 3370–3377 (2006)
12. Cantor, C.R., Schimmel, P.R.: Biophysical Chemistry: Part III: The Behavior of Biological Macromolecules. W. H. Freeman, San Francisco (1980)