Efficient DNA Coding Algorithm for PCR Amplification Information Retrieval Supplementary Materials

For Quick Location Click Here

1. Analysis of Nonspecific Pairing Constraint

Kayama used the recurrent neural network to predict the PCR amplification success rate of specific primer set and DNA template. The experimental results showed that when the DNA sequence had nonspecific pairing with several consecutive positions at the 3 'end of the primer, PCR amplification might start from the nonspecific pairing position, resulting in incomplete amplified sequence[1]. The larger the absolute value of ΔG, the more stable the double strand is[2]. Based on this, the effect of the nonspecific pairing number between the DNA sequence and the 3 'end of the primer on the success rate of PCR amplification was studied.

In synthetic biology, the quantitative indicators to measure the binding stability between two DNA molecules mainly includ Tm and ΔG. Tm value represents the melting temperature, which means the temperature at which the double helix structure of two DNA molecules is destroyed and transformed into a single-stranded DNA molecule (ss-DNA). For DNA sequences with a sequence length of 14nt or less, the Tm value is calculated as follows:

$$Tm = (n(A)+n(T))\cdot 2 + (n(G)+n(C))\cdot 4 \tag{1}$$

Among them, $n(A)$, $n(T)$, $n(C)$ and $n(G)$ represent the number of adenine, thymine, cytosine and guanine in the DNA sequence, respectively. When the length of DNA sequence is greater than 14nt, the above calculation formula will produce a large error. In this case, the correction calculation formula of Tm value is as follows.

$$Tm = 64.9 + 41\times(n(G)+n(C)-16.4)/(n(A)+n(T)+n(G)+n(C)) \tag{2}$$

Gibbs free energy ΔG is an important physical quantity used to measure whether a thermodynamic process can proceed spontaneously and the direction of the reaction. When $\Delta G < 0$, the thermodynamic process can proceed spontaneously. On the contrary, when $\Delta G > 0$, the reaction can not be carried out spontaneously, and the outside world needs to provide energy to the reaction system to make the reaction occur. For the two strands of DNA molecules, two hydrogen bonds can be formed between adenine and thymidine, and three hydrogen bonds can be formed between cytosine and guanine. Therefore, under the principle of base complementary pairing, two single-stranded DNA molecules will spontaneously combine together through hydrogen bonds to form secondary structures. The Gibbs free energy is calculated as follows:

$$\Delta G = \Delta H - T\frac{\Delta H}{T} \tag{3}$$

We designed a large-scale simulation experiment, and selected *Tm* and ΔG (ΔG in the experiment was absolute value) to explore the binding stability between the DNA sequence and the primer when the DNA sequence and the primer 3 'end produced different bits of nonspecific pairing. For 880 primer sequences, we designed DNA sequences in turn, so that at different positions of the DNA sequence, a number of continuous nonspecific pairs with the 3 'end of the primer sequence were generated, and the length of the primers was 20bp. The average *Tm* and ΔG values were calculated, as shown in Figure 1.

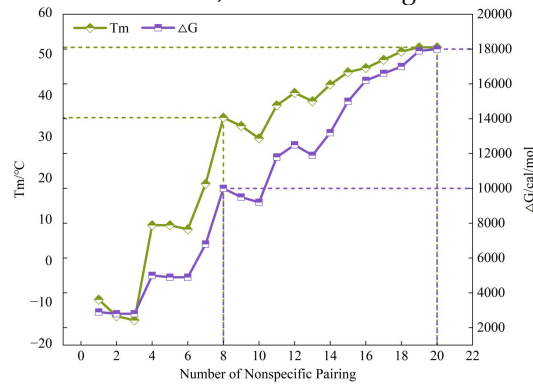**Figure 1.** *Tm* and ΔG values of DNA sequences and primers for different number of nonspecific pairing bits.

The experimental results showed that the more bits of DNA sequence and primer 3 'end produced nonspecific pairing, the higher *Tm* and ΔG values, which represented a more stable binding to the primer and conformed to the PCR amplification law. The *Tm* and ΔG values of DNA sequences and primers were lower when the number of unspecific pairing bits was less than 8 at the 3 'end of DNA sequences and primers, indicating that DNA sequences and primers were not easy to be unspecific pairing at this time, that is, PCR amplification was not error-prone. When the DNA sequence and the primer 3 'end produced a continuous 8 bits of non-specific pairing, the *Tm* and ΔG values will rise sharply, more than 50% of the maximum value, indicating that the primer is more prone to non-specific pairing with the DNA sequence to amplify the interference sequence. When the number of non-specific pairing bits between DNA sequence and primer 3 'end exceeds 8, the combination of DNA sequence and primer is more stable and prone to the risk of amplification interference.

Through the above simulation experiments, this paper found that the complementation of the consecutive 8 bases at the 3 'end of the primer sequence would lead to the Gibbs free energy $\Delta G < -10^4$, $Tm > 35°C$. When the order of magnitude of ΔG exceeds 10^4 and $\Delta G < 0$, The binding stability between the two DNA sequences will affect the selective PCR amplification reaction and interfere with the primer binding site.

In order to explore and confirm that when the coding DNA sequence has a complementary sub-base sequence with the 3 'end of the primer sequence, it will cause a significant adverse effect on the PCR amplification reaction. We designed the DNA sequence for the confirmatory experiment and amplified the sequence through the actual PCR amplification reaction. Then the sequencing reads are obtained by DNA molecular sequencing technology, and finally the final conclusions are determined by analyzing the sequencing reads.

We randomly selected three different primer sequences from the primer library with a GC content of 45% to 60%, as shown in Table 1.

Table 1. Primer sequences selected for the confirmatory experiments.

Primer Number	Primer Sequence	GC Content
P_1	TTCTCCTAGCGCTCTCTAAG	50%
P_2	TGCCACTGATCACGACTGCC	60%

P$_3$	GTTCATAGACTCATCTAGCG	45%

We designed three different sets of DNA sequences for each primer, so that in the template sequence, there were one, two and three segments respectively, which generated sites complementary to the 3 '-end of the corresponding primer sequence with consecutive N bases. Take the 3-segment subsequence as an example, as shown in Figure 2.
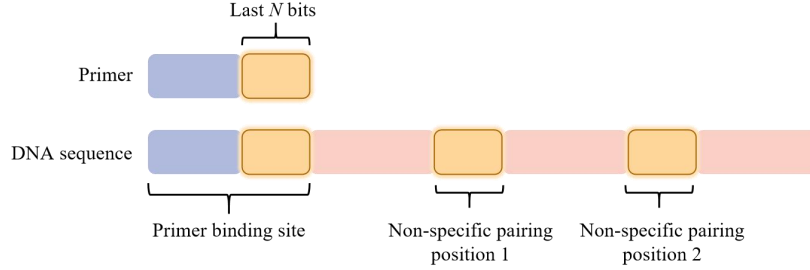

Figure 2. Schematic representation of a DNA sequence and the 3 'end of a primer producing a nonspecific pairing of consecutive N positions.

In a DNA sequence, a primer binding site is a pairing position with a complete and exact match to the primer sequence, and is also the correct position for the specific binding of the primer to the file DNA sequence in a DNA storage system. In general, the correct primer binding site is at the head position in the file DNA sequence. The reason is that the primer sequence will be completely retained in the process of DNA amplification, and from the 3 'end of the primer sequence, the new DNA sequence complementary to the template DNA sequence will be completely copied until the end of the sequence. If the primer sequence is combined with the DNA template sequence at the nonspecific pairing position 1, in the current round of PCR amplification reaction, when the sequence is replicated to obtain a new DNA sequence, it will only start from the nonspecific pairing position 1 until the end of the sequence, and the previous sequence will be lost. The same is true for amplification at nonspecific pairing position 2.

The selective PCR amplification reaction is a cyclic process with multiple rounds of reactions to achieve exponential replication amplification of the target DNA sequence. If in the first round of PCR amplification reaction, 50% of the primers are bound at the nonspecific pairing position 1, only 50% of the sequences are correct in this round of reaction. When the loop enters the second round, the proportion of correct sequences drops to 25%, and the third round drops to 12.5%... . Therefore, there is an exponential decrease in the number of correct sequences, making the access to the target sequence fail.

We designed a total of nine template sequences, whose sequence parameters are shown in Table 2.

Table 2. Template sequence parameters for confirmatory experiments.

Sequence Number	Corresponding Primer Number	Number of Nonspecific Pairing Sites	Total Number of Pairing Sites	Sequence of Nonspecific Pairing Positions
S$_1$	P$_1$	0	1	TCTCTAAG
S$_2$	P$_1$	1	2	TCTCTAAG
S$_3$	P$_1$	2	3	TCTCTAAG
S$_4$	P$_2$	0	1	CGACTGCC
S$_5$	P$_2$	1	2	CGACTGCC
S$_6$	P$_2$	2	3	CGACTGCC

S_7	P_3	0	1	ATCTAGCG
S_8	P_3	1	2	ATCTAGCG
S_9	P_3	2	3	ATCTAGCG

The PCR amplification reaction was performed on the synthesized DNA file pool, where the parameters of the PCR amplification reaction are shown in Table 3.

Table 3. PCR amplification reaction parameters.

parameters	value	parameters	value
10× amplification buffer	$10\ \mu L$	dGTP concentration	$200\ \mu mol\ /\ L$
Primer concentration	50pmol each	Template DNA sequence	$1\ \mu g$
$c(Mg^+)$	1.5mmol	TaqDNA polymerase	$2.5\ \mu$
dATP concentration	$200\ \mu mol\ /\ L$	Volume of reaction liquid	$100\ \mu L$
dTTP concentration	$200\ \mu mol\ /\ L$	Number of PCR amplification cycles	30
dCTP concentration	$200\ \mu mol\ /\ L$		

The DNA molecules of the nine groups of products after the PCR amplification reaction were sequenced, and the sequencing result files in FASTQ format were obtained, whose format is shown in Figure 3.

```
@ST-E00126:128:HJFLHCCXX:2:1101:7405:1133
TTGCAAAAAATTTCTCTCATTCTGTAGGTTGCCTGTTCACTCTGATGATAGTTTGTTTTGG
+
FFKKKFKKFKF<KK<F,AFKKKKK7FFK77<FKK,<F7K,,7AF<FF7FKK7AA,7<FA,,
```

Figure 3. FASTQ files.

Figure 3 shows a basic unit of a sequencing file, called a readout sequence or a read. Each read consists of four lines, and the first line shows the information of the sequencer and the coordinate information of the sequencing readout. The second line begins with the readout for sequencing, which has the full range A, T, C, G, and N. N means that the sequencer cannot determine the specific type of base based on the readout signal, that is, the base here cannot be read. The third row stores the reserved line for additional information. The fourth row stores the read-out quality value of the base at each position of the sequencing read, which is in one-to-one correspondence with the sequencing read sequence in the second row. This line uses the ASCII encoded character value as the pthred value, which is the base quality value. The quality value indicates the sequencing quality and reliability of the base, the higher the value, the more reliable.

The FASTQ files of the three groups of experiments were statistically analyzed, and the experimental results obtained are shown in Table 4. From the experimental results, it is easy to see that as long as there is nonspecific pairing between the DNA sequence and the 3 'end of the primer, the primer will bind to the DNA sequence at the nonspecific pairing position,

thus amplifying invalid short sequences. When the number of unspecific pairing bits was less than 8, the proportion of invalid short sequences was relatively small, ranging from 1.2% to 9.9%, and the success rate of PCR amplification was from 90.1% to 98.8%. However, when the number of nonspecific pairing bits reached 8 or more, the proportion of invalid short sequences increased, and the success rate of PCR amplification significantly decreased to 80.9% or less. It is proved that the non-specific pairing of DNA sequence and primer will produce invalid interference amplification sequences, resulting in a decrease in the success rate of PCR amplification. When the number of non-specific pairing bits is 8 or more, it will have a more serious impact. This experimental result further proves the influence of nonspecific pairing situation in DNA sequence on PCR amplification.

Table 4. Proportion of each amplified sequence in different cases of non-specific pairing number.

Amplified Sequence Type	Percentage(%)									
	$N=2$	3	4	5	6	7	8	9	10	11
1 2 3	98.8	97.9	95.2	93.7	91.5	90.1	80.9	77.1	72.7	69.8
2 3	1.2	2.1	4.3	4.5	6.2	7.0	10.7	12.8	14.9	16.4
3	0	0	0.5	1.8	2.3	2.9	8.4	10.1	12.4	13.8

In summary, in order to ensure the specificity of the PCR amplification reaction and improve the efficiency of PCR amplification, the generated DNA sequence should try to avoid the nonspecific pairing of consecutive 8 bases with the 3 'end of the primer (hereinafter collectively referred to as "nonspecific pairing constraint"). Based on this constraint, this paper proposes a novel efficient DNA coding algorithm for information retrieval of PCR amplification.

References

1. Karami, A.; Gill, P.; Kalantar Motamedi, M.; Saghafinia, M. A review of the current isothermal amplification techniques: applications, advantages and disadvantages. Journal of global infectious diseases 2011, 3, 293-302.

2. Mann, T.; Humbert, R.; Dorschner, M.; Stamatoyannopoulos, J.; Noble, W.S. A thermodynamic approach to PCR primer design. Nucleic acids research 2009, 37, e95-e95.

2. Analysis of Primer Statistical Properties

Upon further investigation, it was observed that the base distribution of primer sequences in the primer library exhibits specific statistical characteristics and is not uniformly distributed. For instance, base combination lengths of 4 and 5 nucleotides demonstrate varying frequencies, as illustrated in Figure 4. Some base combinations appear with high frequency, while others appear with low frequency, due to primer design guidelines rather than random generation.
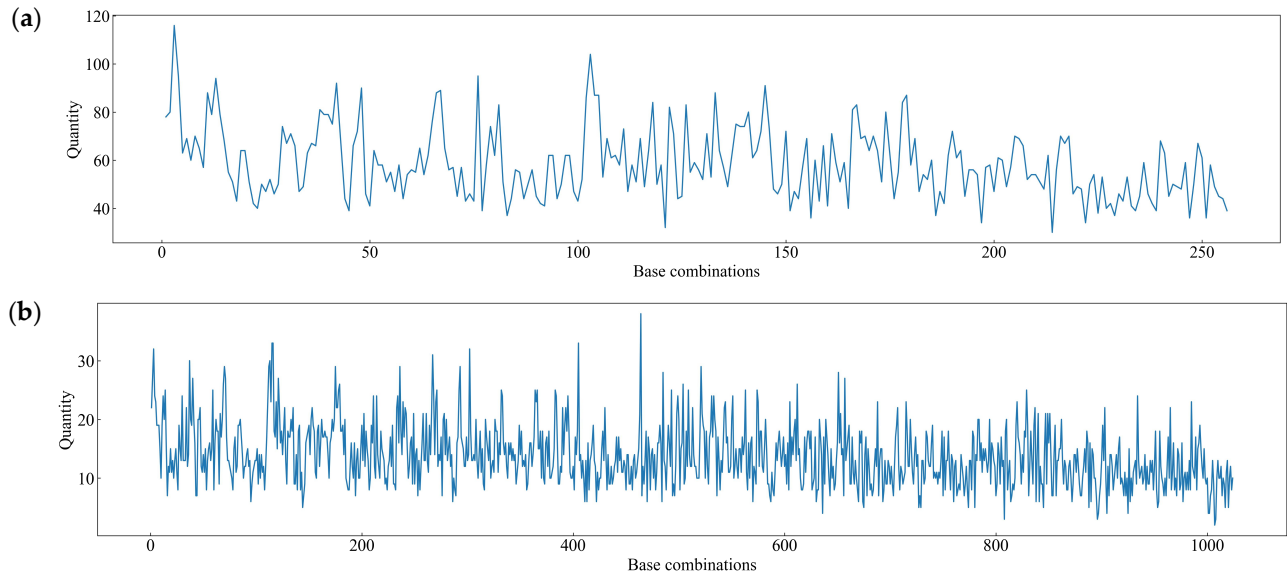
(a)

(b)

Figure 4. Statistical properties of primer libraries: (a) represents number of base combinations of length 4 (256 in total) and their frequency in the primer library; (b) represents number of base combinations of length 5 (1024 in total) and their frequency in the primer library.