STARP Primer Design

Kaleb Burnham

USDA ARS

Fargo, North Dakota

Abstract

Dr. Yunming Long introduced Semi-thermal Asymmetric Reverse PCR (STARP) in 2016. The objective of the Starp Primer Design application here described is to simplify the process of creating primers for use in this technique. Four main stages compose the program: parsing single nucleotide polymorphisms (SNPs), the creation of potentially useful reverse primers, the creation of two asymmetrically modified allele specific (AMAS) primers, and the combination of reverse primers and AMAS primer.

1. **Technology**
2. **Program Input**

There are three acceptable formats a user may enter data to the program.

**SNP Sequence**

In a SNP sequence, SNPs are denoted by brackets. For example, [G/C] represents a substitution SNP while [A/-] represents a deletion and [-/T] represents an insertion. In general, a SNP may be [x/y] where x and y are elements of {A, C, T, G, -} and x **≠** y. A full sequence might look like the following:

CCCGTCTTCTACACCCGTCTCTGTGGGCCC[A/G]TTCAAGTGCTGGGAGGGCACCCTCCAAGATCGACTTTAC[G/-]GACTGCTCAAGGTCTGGCGGTATGGTGGTATGGCGAGCCCGTACCGTGATACATGCGGCA

To recognize this format, the parsers look for the [x/y] pattern.

**Two Alleles**

Rather than including all SNPs in one sequence, one can compare two equal-length sequences. Here, the program expects four lines: the first and third lines are FASTA headers such that they start with a ‘>’ character, and the second and fourth lines are the DNA sequences. For example, the following snippet has a deletion at the 23rd position and a T > A substitution at the 38th position.

>Allele 1

CCCTACGCCCGTCTTCTACACCCGTCTCTGTGGGCCCTGTTCAAGTGCTG

>Allele 2

CCCTACGCCCGTCTTCTACACC-GTCTCTGTGGGCCCAGTTCAAGTGCTG

To recognize this format, the parsers look for a ‘>’ symbol at the beginning of the first and third lines and equal length sequences following them.

**Single Blast Result**

In a single blast result, one may directly paste results from a BLAST query. The parser tokenizes the input by splitting it on whitespace and looks for the ‘Query’ and ‘Sbjct’ lines. An example of valid input is the following.

Query 2643 TATCTTCATTGTATTGATTTTATAACCGATTCCAAAATGTATTCTTAAAGGTACATCATC 2702

|||| ||||||||||||| | |||| ||| ||||| |||| ||||||||| |||||

Sbjct 586624978 TATCCTCATTGTATTGATCTATTAACTAATTATTAAATGCATTCATAAAGGTACCTCATC 586625037

Query 2703 GTAATTGATGATATATGGGATGAAAAAGTGTGGGAATTCATTAA-T-TGCGCTTTCTCCA 2760

||||| |||||||||||| ||||||||| ||||| || |||| | ||| |||| ||||

Sbjct 586625038 GTAATCGATGATATATGGAATGAAAAAGCATGGGAGTTACTTAAGTGTGC-CTTT-TCCA 586625095

To recognize this format, the parsers look for a ‘|’ character.

1. **Reverse Primer Generation**

The reverse primer generation method is a brute-force technique on the 5’->3’ strand. All subsequences common to both alleles of length 18 to 27 are recorded as potential primers. Then, the sequences that meet any of the following conditions are removed.

1. Contains any non-nucleotide character.
2. Contains 10+ contiguous G/C characters or 12+ A/T characters.
3. Contains 8+ of any single nucleotide.
4. Contains 6+ dinucleotide repeats.
   1. Eg, ‘ACACACACAC’ contains a 5-repeat of ‘AC’
5. Contains GC content < 20% or > 80%.
6. Has melting temperature < 53 °C or > 62 °C.

Following this filter, the remaining primers are filtered based on their binding sites on the alleles and nontarget sequences. Any primers with multiple binding sites are removed.

**Definition 1:** A **binding site** of a primer or its reverse complement is a subsequence that has less than five mismatched nucleotides, and the primer’s 5’ nucleotide is matched or there are fewer than 2 mismatches at the 2nd, 3rd, and 4th positions from the 5’ end, or the reverse complement’s 3’ end is matched or there are fewer than 2 mismatches at the 2nd, 3rd, and 4th positions from the 3’ end.

For the primers with melting temperature between 53 °C and 58 °C, a cartesian product is performed with its sequence and the set {‘’, ‘C’, ‘G’, ‘CG’, ‘GC’, ‘CGC’, ‘GCG’} and the results are added as potential primers. Then, primers are filtered once again and removed if they meet any of the following criteria.

1. Contains 28+ nucleotide bases.
2. Has melting temperature > 62 °C.
3. Contains 10+ contiguous self-complementarity
4. Primer length – self-complementarity <= 4

After filtering is complete, the primers are sorted according to the following multi-key sort. Note that the phrasing indicates “bad” primers evaluate to True, so they are pushed towards the end of the list.

1. Contains 9 contiguous G/C or 11 contiguous A/T.
2. Contains a mononucleotide repeat of length 7+.
3. Contains contain a dinucleotide repeat of length 5+.
4. GC content is < 25% or > 75%.
5. Contiguous complementary score >= 8.
6. Primer length – complementary score <= 6.
7. Contains 8 contiguous G/C or 9 contiguous A/T.
8. Contains a mononucleotide repeat of length 6+.
9. Contains a dinucleotide repeat of length 4+.
10. GC content is < 30% > 70%.
11. Contiguous complementary score >= 7.
12. Primer length – complementary score <= 8.
13. Contains 6 contiguous G/C or 7 contiguous A/T.
14. Contains a mononucleotide repeat of length 5+.
15. Contains a dinucleotide repeat of length 3+.
16. GC content is < 35% or > 65%.
17. Contiguous complementary score >= 6.
18. Primer length – complementary score < 10.
19. Contains 6 A/T or 5 G/C in the first seven bases.
20. Contains 4 A/T or 3 G/C in the first four bases.
21. Contains a mononucleotide repeat of length 4+
22. GC content is < 40% or > 60%.
23. Contiguous complementary score >= 4.
24. Primer length – complementary score <= 14.
25. Length closest to 22 nucleotides.
26. **AMAS Primer Creation**

The application attempts to create AMAS primers for each SNP both upstream and downstream, though not all SNPs will have acceptable primers. The method for generating these differs between substitution and indel SNPs.

**Substitution SNPs**

Potential upstream primers are those in the set of sequences from both alleles of length 16-25 with the last nucleotide being at the SNP position. Primers of equal length but different alleles constitute a pair.

Potential downstream primers are those in the set of sequences from both alleles of length 16-25 with the first nucleotide being the SNP position. These are paired similarly.

These pairs are filtered out if one of their primers meet one of the following criteria.

1. Contains 10+ contiguous G/C nucleotides or 12+ contiguous A/T nucleotides.
2. Contains a mononucleotide length of 8+.
3. Contains a dinucleotide length of 6+.
4. Has GC content < 20% or > 80%.

If one exists, the pair where both primers have melting temperature between 53 °C and 60 °C and the average is closest to 58 °C is chosen to proceed forward (this step is performed separately for upstream and downstream pairs). The primers in this pair will undergo a nucleotide substitution as described by Dr. Long in Appendix I.

**Indel SNPs**

Indel primers are different in that they overlap the SNP position. For upstream primers, they are the set of primer pairs of length 17 to 26 where the first 16 positions come before the SNP. Then, any pairs with the same nucleotide at the 3’ end are removed. The remaining list is sorted based on the number of differences in the last four bases (more being preferable) and primer length (shorter is preferable). Select the first pair to proceed.

For downstream primers, they are the set of sequences of length 17 to 26 where the first 16 positions come after the SNP. Then, any pairs with the same nucleotide at the 5’ end are removed. The remaining list is sorted based on the number of differences in the first four bases (more being preferable) and primer length (shorter is preferable). Select the first pair to proceed. Since no pair can contain an indel between the two primers, this effectively converts indels into substitution SNPs. Section 5 and Appendix I use this fact.

The primers in each pair will also undergo a nucleotide substitution described in Appendix I.

1. **Combination of AMAS and Reverse Primers**

At this stage, there is a list of reverse primers and possibly an upstream and downstream primer pair for each SNP. The set of reverse primers that can be associated with each primer pair depend on the following conditions. By default, the first three primers that meet these conditions will be for each primer. These reverse primers are still sorted so there is no need to do that again.

1. They come must come after upstream primers and must come before downstream primers.
2. Length of primer – complementary score between primer and AMAS primer > 5 for both AMAS primers.

Now, a cartesian product is performed on each primer pair and its corresponding reverse primers. Each reverse primer determines what tail needs to be added to the AMAS primers. This process will be described next.

**The Addition of Tails to AMAS Primers (Incomplete)**

As defined by Long et al., the two possible tails are

Tail 1: GCAACAGGAACCAGCTATGAC

Tail 2: GACGCAAGTGAGCAGTATGAC

For each tail/primer combination, trim off the (maximum 6) nucleotides that overlap with the primer. For example, using Tail 1 and primer GACTACGCGCATCT, the resulting concatenation is

GCAACAGGAACCAGCTATGAC + GACTACGCGCATCT = GCAACAGGAACCAGCTATGACTACGCGCATCT.

Then, define Amplicon 1 and Amplicon 2.

Amplicon 1: The length of PCR product in the first allele.

Amplicon 2: The length of PCR product in the second allele.

The following rules determine which tail is added to each AMAS primer.

1. If Amplicon 1- Amplicon 2 ≥ 8,

Tail 1 is concatenated to the first AMAS primer.

Tail 2 is concatenated to the second AMAS primer.

1. If 1≤ Amplicon 1- Amplicon 2 ≤ 7,

Tail 2 is concatenated to the first AMAS primer.

Tail 1 is concatenated to the second AMAS primer.

1. If Amplicon 1- Amplicon 2 = 0,

Tail 1 or 2 will be removed the overlapped bases at 3ʹ end and then add to F1 or F2 primer following the principle in the paper;

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **SNP type** | **AS-nucleotide** | **Designated allele name** | **Assigned tail\*** | **Designated AMAS-primer name** | **Nucleotide position from the 3′ end for substitution** |
| [C/G] | C | Allele 1 | [Tail 1] | AMAS-primer 1 | 4th |
|  | G | Allele 2 | [Tail 2] | AMAS-primer 2 | 3rd |
| [C/T] | C | Allele 1 | [Tail 1] | AMAS-primer 1 | 4th |
|  | T | Allele 2 | [Tail 2] | AMAS-primer 2 | 3rd |
| [C/A] | C | Allele 1 | [Tail 1] | AMAS-primer 1 | 3rd |
|  | A | Allele 2 | [Tail 2] | AMAS-primer 2 | 4th |
| [G/T] | G | Allele 1 | [Tail 1] | AMAS-primer 1 | 4th |
|  | T | Allele 2 | [Tail 2] | AMAS-primer 2 | 3rd |
| [G/A] | G | Allele 1 | [Tail 1] | AMAS-primer 1 | 3rd |
|  | A | Allele 2 | [Tail 2] | AMAS-primer 2 | 4th |
| [T/A] | T | Allele 1 | [Tail 1] | AMAS-primer 1 | 3rd |
|  | A | Allele 2 | [Tail 2] | AMAS-primer 2 | 4th |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |  |
| T | Allele 2 | [Tail 2] | AMAS-primer 2 | 3rd |
| [G/A] | G | Allele 1 | [Tail 1] | AMAS-primer 1 | 3rd |
| A | Allele 2 | [Tail 2] | AMAS-primer 2 | 4th |
| [T/A] | T | Allele 1 | [Tail 1] | AMAS-primer 1 | 3rd |
| A | Allele 2 | [Tail 2] | AMAS-primer 2 | 4th |

1. If -7≤ Amplicon 1- Amplicon 2 ≤ -1,

Tail 1 is concatenated to the first AMAS primer.

Tail 2 is concatenated to the second AMAS primer.

1. If Amplicon 1- Amplicon 2 ≤ -7,

Tail 2 is concatenated to the first AMAS primer.

Tail 1 is concatenated to the second AMAS primer.

**Appendix I - Substitution Rules**

These have been changed since this document was first created. See ‘STARP F primer design\_clarified.docx’.