A Genome-Specific PCR Primer Design Program for Open Reading Frames

Kwoh Chee Keong[1], Kok Wui Lim [2]
[1] *School of Computer Engineering, Nanyang Technological University,* Singapore
[2]*Genome Institute of Singapore, Singapore*
Email : asckkwoh@ntu.edu.sg

ABSTRACT Proper PCR primer design determines the success or failure of Polymerase Chain Reaction (PCR) reactions. In this project, we develop GENE-PRIMER, a genomes specific PCR primer design program that is amenable to a genome-wide scale. To achieve this, we incorporated various parameters with biological significance into our program, namely, primer length, melting temperature of primers Tm, guanine/cytosine (GC) content of primer, homopolymeric runs in primer and self-hybridization tendency of primer. In addition, BLAST algorithm is utilized for the purpose of primer specificity check. In summary, selected primers adhered to both physico-chemical criteria and also display specificity to intended binding site in the genome.

Keywords: PCR, Primer, BLAST, Open Reading Frame, DNA.

1 INTRODUCTION

The discovery of Polymerase Chain Reaction (PCR) by Kary Mullis has revolutionized the field of molecular biology. Basically, PCR is a molecular biology technique that enables the amplification of a single copy of target DNA sequence in an exponential manner [4]

Molecular biologists studying the functional aspect of a particular gene of interest can use PCR to isolate and generate large amount of the target DNA before proceeding to further investigative techniques such as DNA mutagenesis or DNA sequencing [3].

In the field of population genetics, PCR-based methods such as Arbitrary Primer PCR (AP-PCR) have been used to gauge the extent of genetic variations in populations of organisms. In essence, PCR has really revolutionized the approach to biological research.

2 MOTIVATION & APPROACH

One of the most important factors for successful PCR is proper primer design. The chosen pair of primers must not only match the flanking regions of the target DNA, they must also adhere to certain criteria such as primer specificities, guanine/cytosine (GC) contents and melting temperatures (Tm). Although primer design can be done manually, it is a laborious and time-consuming process. This coupled with the huge amount of sequence data that are available, makes manual primer design nearly an impossible task.

The main objective of this project is (1) to develop a genome-specific PCR primer design program. (2) automate primer design on a genome-wide scale and have the capability to design primers for multiple inputs. (3) The primer design program must incorporate the various requirements for good PCR primer design.

All these parameters are listed below.

i. Primer length between 20 - 30 base pairs.
ii. GC content of 40 - 60%.
iii. Homolopolymeric run in primer ranges from 1 - 5 base pairs.
iv. Minimum self-hybridization in the form of primer-dimers.
v. Melting temperature in the range of 50 - 690C.
vi. Site-specificity of primer meaning that primers should not falsely anneal to other sites except the intended site.

In this project, we are given a genome containing a set of Open Reading Frames (ORFs), $G = \{g_1, g_2, \ldots, g_p\}$. We would like to find pairs of primers for amplifying known genes (ORFs) from the complete genome with the additional constraint that these primers must satisfy most, if not all, of the requirements defined by the user. In addition, the global primer specificity constraint ensures that the amplified region can uniquely identify the ORF, g_i, from which it originated from.

3 IMPLEMENTATION

3.1 System Requirements

DNA sequences representing individual Open Reading Frames (ORFs) of the completed the *Schizosaccharomyces pombe* (fission yeast) genome were downloaded in FASTA format from the Genbank database at the National Center for Biotechnology Information (NCBI) (*http://www.ncbi.nlm.nih.gov/entrez*). These downloaded DNA sequences serve two purposes. They are used as test sequences and also for the formatting of genome-specific BLAST database. In total, we downloaded 5006 DNA sequences representing all the ORFs of the genome.

We implemented the primer design program on a terminal with SSH Secure Shell connection to a Sun Microsystems's Ultra® Sparc II server running Solaris® operating system. Formatting of the genome-specific BLAST database was done locally. All programs were written in Perl programming language (Version 5.6.1).

3.2 Implementation Details

The primer design program is implemented in two phases, namely the initial primer generation phase and the specificity filtering phase. The primer generation phase will generate lists of candidate forward and reverse primers that conform to user-specified parameters. In the specificity filtering step, we first query the output from the primer generation phase against pre-formatted local BLAST database containing all the DNA sequences of *S.pombe* genome employing the BLASTN program of the Basic Local Alignment Search Tool (BLAST) family of programs [1]. Non-specific candidate primers that show sequence similarities to multiple sites in the genome are discarded. Figure 1 illustrates the flow of the primer design program.

Primer generation phase

Basically, primers are strings over the DNA alphabet, $\sum_{DNA}=\{A,T,G,C\}$ with the set of all these strings being $\sum^*$. In the primer generation phase, candidate primers with user-defined length are generated recursively from Open Reading Frames (ORFs). The forward primers are derived recursively (substrings) from within a defined window, w on the 5' end of the plus (+) strand of DNA. Similarly, the reverse primers are derived from the window on the 5' end of minus (-) strand of DNA. See Figure 2.

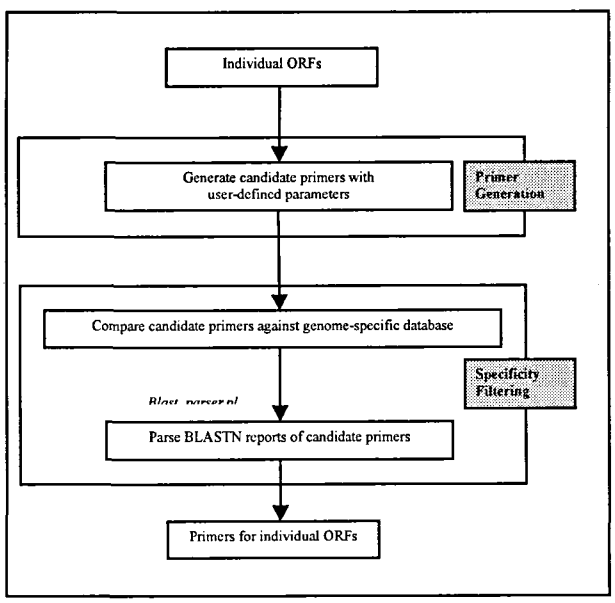

Figure 1: *Overview of implementation.*

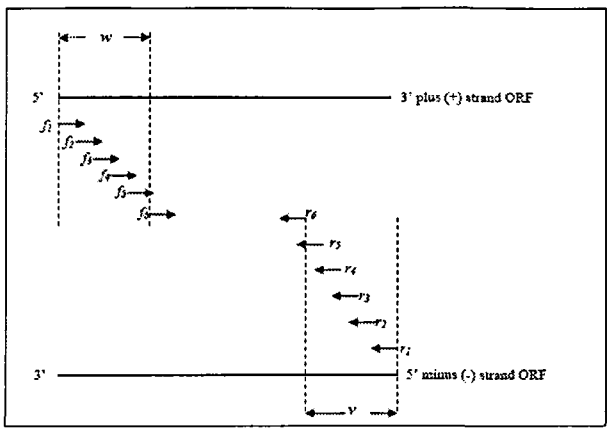

Figure 2: *Forward primers,* $f=(f_1,f_2,.....f_n)$ *and reverse primers,* $r=(r_1,r_2,.....r_n)$ *derived from windows, w and v respectively.*

By default, the defined windows, w and v, constitute approximately 10% of an ORF on the 5' end and 3' end respectively. The starting positions for all candidate primers within a defined window will be different from one another. This means that f_1 starts from position p_i, f_2 starts from position p_{i+1} , and f_n starts from position p_{i+w-1}, where w is the window length. As such, the number of candidate primers will be proportional to the window length, w. Hence, we can expect the generation of w candidate primers from each direction.

Computation of various user-defined parameters namely, primer length, melting temperature (T_m), CG content, maximum number of contiguous bases in primer (homopolymer runs) and the number of self-complementary bases in primer are performed in this phase.

For calculation of primer melting temperature (^oC), we use the following formula in the subroutine *calculateTm*.

$$T_m\,(^oC) = 4\,(G+C) + 2\,(A+T)$$

where G is the number of guanine nucleotides
C is the number of cytosine nucleotides
A is the number of adenine nucleotides
T is the number of thymine nucleotides

We apply the following formula for the calculation of CG content in subroutine *calculateCG*.

$$\text{CG}\,(\%) = \left[\frac{G+C}{l}\right] \text{x } 100$$

where G is the number of guanine nucleotides in primer
C is the number of cytosine nucleotides in primer
l is the primer length

In subroutine *computemaxContigBase* of our program, we calculate the maximum of homopolymeric regions in candidate primers. Homopolymeric regions are regions containing stretches of the same nucleotide. For each

candidate primer, starting from base 1, we compare pairs of adjacent bases iteratively. If bases b_i and b_{i+1} are the same, we increment the count of contiguous bases whereas if they are not the same, we revert to the default count which is 1. For example, candidate primer 5'-ATTTTCGTT-3' will give the counts (1 2 3 4 1 1 1 2 1) when comparing pairs of adjacent bases iteratively. The maximum value in the array is the maximum of number of homopolymer in primer.

Maximum contiguous base = max $\{C(b_i, b_{i+1})\}$

$C(b_i, b_{i+1})$ is the count of contiguous bases in a pair of adjacent nucleotides.

Next, we model the tendency of self-hybridization in candidate primers. As shown in Figure 3, self-hybridization occurs due to complementary base pairing in overlapping region of the 5' end candidate primer, p, and the 5' end of its reverse, p'.

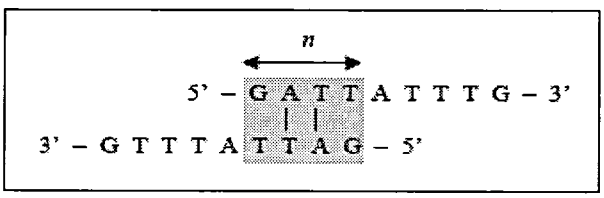

Figure 3: *Self-hybridization occurs in the overlapping region (shaded area) of primer, p and its reverse, p'.*

To determine the maximum number of inter-strand bonds, we need to compute the number of inter-strand bonds for each pair of 5' end primer and 5' end of its reverse varying in length of the overlapping region, $n = (1, 2, ...l)$. We propose a scoring scheme $s(b_i, b_j)$ for scoring the pairing of nucleotides in the primer strand, b_i and the nucleotides in its reverse strand, b_j.

$$s(b_i, b_j) = \begin{cases} 1 & if \{b_i, b_j\} = \{A, T\} \\ 1 & if \{b_i, b_j\} = \{C, G\} \\ 0 & else \end{cases}$$

Subsequently, we compile all the scores, $s(b_i, b_j)$, of for each pair of 5' end primer and 5' end of its reverse into $Hyb(p, p')$. All the possibilities of self-hybridization for a candidate primer are shown in Figure 4. $Hyb(p, p')$ measures the total number of inter-strand bonds in overlapping region of length n.

$$Hyb(p, p') = \sum_{i=1}^{n} s(b_i, b_j)$$

where n is the length overlapping region.

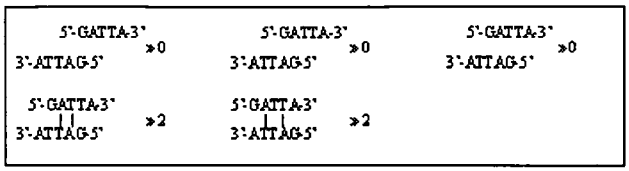

Figure 4: *Self-hybridization of hypothetical candidate primer*

Lastly, we take the maximum of $Hyb(p, p')$ as the final measure of tendency for a candidate primer.

$$MaxHyb(p, p') = \underset{n=1,2,...l}{Max} \{Hyb(p, p')\}$$

OR

$$MaxHyb(p, p') = \underset{n=1,2,...l}{Max} \left\{ \sum_{i=1}^{n} s(b_i, b_j) \right\}$$

Specificity filtering phase

The motivation for incorporating this phase in our primer design program is that the ideal primers should only bind to intended specific sites on the genomic DNA template. One of the measures to ensure specificity of primers is to design primers with length ranging from 20-30 nucleotides as the longer the primer sequences are, the higher the probability that they are unique and therefore only bind to designated sites [3].

In addition, we would like to use BLASTN program to map PCR primers to the genome based on sequence similarities [2]. The idea here is to use sequence similarities to differentiate between specific and non-specific primers.

We define non-specific primers as primers that show high sequence homology or sequence similarities to multiple ORFs in the database. On the other hand, specific primers are primers that show low similarities to other database hits and hence represent unique primers. More specifically, candidate primers that are similar (percent identity ~ 100%) and aligned entirely (hit length same as primer length) to many ORFs (more than 2 ORFs) are considered as not specific

We used the default scoring scheme (+1/-3 for match and mismatch respectively) for the BLASTN searches as this scoring scheme favors near identity between the candidate primers and the genome. These settings are optimized for mapping PCR primers to the genome with high statistical significance [2].

4 DISCUSSION

4.1 From Open Reading Frames (ORFs) To PCR Primers

We have successfully implemented a PCR primer design program to process Open Reading Frames (ORFs) in FASTA format into list of PCR primers. The transition involves manipulations of DNA strings which is made

possible by the array of powerful text processing functionalities offered by Perl programming language. In addition, portability of the program is not an issue as programs written in Perl work in the Unix environment as well as in the Windows environment. As many bioinformatics programs are written in Perl, this effort is in line with the current trend in the field of bioinformatics. Primer design programs will not be complete without incorporating various parameters with underlying biological significance. This is exemplified by our effort to model self-hybridization tendency in primers. The ability to select primers *in silico* will save valuable time and unnecessary frustrations resulted from failed PCR experiments due to self-hybridization of primers in reaction tubes.

4.2 Comparison With Other Primer Design Program

The only way to assess the performance of primer design programs is to conduct PCR experiments using the designed primers. We report here a comparison with a popular primer design program, Primer3 [5]. Five ORFs were randomly selected as test sequences. We then designed forward primers for these ORFs using both our program and the program Primer3. The primer design parameters are given in Table 1.

Parameter	Value
Minimum length	20
Maximum length	22
Minimum melting temperature (°C)	40
Maximum melting temperature (°C)	65
Minimum CG content (%)	40
Maximum CG content (%)	60
Maximum homopolymeric runs	8
Maximum self-hybridization	10

Table 1: *Primer design parameters used in evaluation.*

When compare to our implementation, Primer3 generated many non-specific primers generated by Primer3 which is detected by our implementation (Refer to Table 2). These non-specific primers represent a major short-coming of their model. This could be due to the fact that Primer3 neglect to check primers for specificity in term of mispriming potential to the genome sequence. It only compares generated primers to some preprocessed libraries of frequently encountered repetitive sequences or microsatellites in genomes of human, mouse and primates which act as specificity check.

ORF	Primer Sequence	Comment
p2003_100401-SPCC794.05c	[illegible]	9 hits to genome database
p2003_100408-SPCC553.05	[illegible]	4 hits to genome database
pJ5566_101203-SPBPB2B2.16c	[illegible]	3 hits to genome database
p4C9_100088-SPAC750.02c	[illegible]	3 hits to genome database
p4C9_100107-SPAC13D1.02c	[illegible]	13 hits to genome database

Table 2: *Non-specific forward primers generated by Primer3.*

4.3 Possible applications

Our primer design program is targeted for the design of PCR primers for large scale amplification of gene-specific probes. These probes can be spotted on cDNA microarrays used in gene expression study. This made possible by the ability of our program to perform primers design in batch processing mode. Another possible application of our program is to design primers for single-target PCR which is routinely performed in many molecular biology laboratories. The advantage offered by our program is that we combine both PCR primer design for single and microarray probes design into a single program.

5 CONCLUSION

This project developed a genome-specific PCR primer design program. Unlike other programs, our program GENE-PRIMER is able to generate primers that are specific to the genome of an organism. This is conferred by the ability to format our own BLAST database which in turn will help users in customizing their primer design needs.

Our program also supports primer design for multiple inputs. This effectively has enabled a genome-wide approach in PCR primer design while incorporated the various parameters with biological significance into our program, namely, primer length, melting temperature of primers Tm, guanine/cytosine (GC) content of primer, homopolymeric runs in primer and self-hybridization tendency of primer. It is worth mentioning that we have also incorporated the BLAST algorithm for the purpose of primer specificity check. In summary, our program will be able to select primers that adhere to both physico-chemical criteria and also display specificity to intended binding site in the genome.

REFERENCES

[1] Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* 25(17): 3389 – 3402.

[2] Korf, I., Yandell, M., and Bedell, J. 2003. *BLAST.* Sebastopol: O'Reilly & Associates Inc.

[3] McPherson, M.J., and Moller, S.G. 2000. *PCR.* Oxford: BIOS Scientific Publishers.

[4] Mullis, K.B., and Faloona, A.F. 1987. Specific synthesis of DNA in vitro via a polymerase-catalysed chain reaction. *Methods in Enzymology* 155: 335 – 351.

[5] Rozen, S., and Skaletsky, H. 2000. Primer3 on the WWW for general users and for biologist programmers. In Krawetz, S., and Misener, S. (eds.), *Bioinformatics Methods and Protocols: Methods in Molecular Biology*. New Jersey: Humana Press, 81 – 88