MFEprimer-3.0: quality control for PCR primers

Kun Wang[1,2,†], Haiwei Li[2,†], Yue Xu[2], Qianzhi Shao[2], Jianming Yi[2], Ruichao Wang[2], Wanshi Cai[2], Xingyi Hang[2], Chenggang Zhang[3], Haoyang Cai[1,*] and Wubin Qu[2,*]

[1]Center of Growth, Metabolism, and Aging, Key Laboratory of Bio-Resources and Eco-Environment, College of Life Sciences, Sichuan University, Chengdu 610064, China, [2]iGeneTech Bioscience Co., Ltd, Beijing 102206, China and [3]Institute of Radiation Medicine, Academy of Military Medical Sciences, Academy of Military Sciences, Beijing 100850, China

ABSTRACT

Quality control (QC) for lab-designed primers is crucial for the success of a polymerase chain reaction (PCR). Here, we present MFEprimer-3.0, a functional primer quality control program for checking non-specific amplicons, dimers, hairpins and other parameters. The new features of the current version include: (i) more sensitive binding site search using the updated *k-mer* algorithm that allows mismatches within the *k-mer*, except for the first base at the 3′ end. The binding sites of each primer with a stable 3′ end are listed in the output; (ii) new algorithms for rapidly identifying self-dimers, cross-dimers and hairpins; (iii) the command-line version, which has an added option of JSON output to enhance the versatility of MFEprimer by acting as a QC step in the 'primer design → quality control → redesign' pipeline; (iv) a function for checking whether the binding sites contain single nucleotide polymorphisms (SNPs), which will affect the consistency of binding efficiency among different samples. In summary, MFEprimer-3.0 is updated with the well-tested PCR primer QC program and it can be integrated into various PCR primer design applications as a QC module. The MFEprimer-3.0 server is freely accessible without any login requirement at: https://mfeprimer3.igenetech.com/ and https://www.mfeprimer.com/. The source code for the command-line version is available upon request.

INTRODUCTION

Primer design strategies vary with the type of applications of a polymerase chain reaction (PCR). For instance, multiplex PCR requires compatible primer sets without non-specific amplicons (1–3). Primer design for the gene encoding 16S ribosomal RNA (16S rRNA) requires primers that cover as many bacterial 16S sequences as possible (4). Another similar example is designing of primers for amplifying the rearranged variable (V) regions of antigen receptors (5). Although the primer design strategies differ, they share the same primer quality control (QC) parameters, such as prevention of undesired amplicons, dimers, and hairpins, and so on. However, most currently available tools do not contain a complete QC procedure. GenomeTester (6), Primer-BLAST (7), In-Silicon PCR (http://rohsdb.cmb.usc.edu/GBshape/cgi-bin/hgPcr), and previous versions of MFEprimer (8,9) focus on evaluation of primer specificity. AutoDimer (10) (https://strbase.nist.gov/NIJ/AutoDimer.htm) focuses on primer dimer and hairpin screening. SNPCheck (https://secure.ngrl.org.uk/SNPCheck/snpcheck.htm) focuses on checking single nucleotide polymorphisms (SNPs) in predicted primer binding regions. In contrast to post-specific analysis tools such as In-Silicon PCR and MFEprimer, Primer3 (11) analyzes non-specific binding sites prior to primer design using species-specific mispriming libraries or build-in search for genome-wide non-specific binding sites with the whole genome as the template. Primer3_masker (12) accelerates this genome-wide search process using a *k-mer* based algorithm. This 'prior-to' strategy is important for large eukaryotic genomes, where the probability of primers binding to repeat regions is high and should be avoided. In this study, we introduced MFEprimer-3.0, a functional and independent primer QC program. The new features of the current release include: (i) improved sensitivity with which specificity can be evaluated using an updated *k-mer* algorithm, which allows mismatches within the *k-mer* except at the first base at the 3′ end; (ii) new algorithms to rapidly identify self-dimers, cross-dimers and hairpins; (iii) the command-line version, which has an added option of JSON (https://www.json.org/) output to enhance the versatility of MFEprimer by acting as a QC module in the 'primer design → quality control → redesign' pipeline; (iv) a function for checking

*To whom correspondence should be addressed. Tel/Fax: +86 10 8914 6623; Email: quwubin@gmail.com
Correspondence may also be addressed to Haoyang Cai. Tel: +86 28 8541 8843; Fax: +86 28 8541 2571; Email: haoyang.cai@scu.edu.cn
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

whether the binding sites contain SNPs, which will affect the binding efficiency among different samples; (v) codes rewritten in Go language (https://golang.org/) to ensure reliability, speed and user-friendliness across multiple platforms.

THE MFEPRIMER-3.0 ALGORITHM

The code of MFEprimer-3.0 was written in Go language, which has concurrency features and utilizes multiple core CPUs. In MFEprimer-2.0, we introduced the k-mer index and search algorithm, which enabled rapid identification of primer binding sites. The limitation is that no mismatch is allowed within the k-mer, which is not true in real PCR [13]. Here, a mismatch within k-mer indicates that when searching for primer binding sites, mismatches are allowed between the 3′ end of the primer subsequence (k-mer) and its binding site sequence. In the current version, the k-mer search algorithm was updated to allow mismatches within k-mer. Furthermore, two new modules were developed to detect primer dimers and hairpins.

The *k*-mer mismatch search algorithm

The core idea of k-mer mismatch search algorithm is to convert mismatch search to match search by generating all possible k-mers for a specific k-mer. Here, we set $k = 9$ as an example. The main steps are as follows: (i) the k-mer index algorithm stores the positions of all 9-mers into a file database; (ii) if one mismatch of its binding sites is allowed for any 9-mer, three possible binding site sequences emerge when the first base is a mismatch. For example, for ATCGATCGA, [T/C/G]TCGATCGA represents the three combinations with first base mismatch. Therefore, theoretically, there are $9*3 = 27$ possible mismatch binding sites for any 9-mer, and 28 sequences including the match binding site; (iii) for any primer, its 3′ end subsequence (here a 9-mer) is first converted to 28-mers. Then the information of these 28 positions is directly retrieved from the database. It is noteworthy that MFEprimer-3.0 does not allow a mismatch at the last base of the 3′ end, which is suitable for most applications.

Dimer and hairpin detection algorithm

The algorithm for detecting dimers and hairpins was modified from AutoDimer [10]. Briefly, for dimer detection: (i) two sequences were incrementally over-lapped by one base in each step; (ii) for each step, the alignment score was calculated and if the score was equal to or greater than the cut-off value, the alignment form (dimer) of this step is stored; (iii) the dimer with the maximum score was sorted. A variation of this algorithm was used to detect hairpins. In addition to the alignment score, the Gibbs free energy [14] of each dimer and hairpin were also calculated.

THE MFEPRIMER-3.0 SERVER

The MFEprimer-3.0 server is available at https://mfeprimer3.igenetech.com/ and https://www.mfeprimer.com/. The website is accessible free of charge and without login requirement.

Input

Primer sequence is the only mandatory input in MFEprimer-3.0. Multiplex PCR primers in FASTA format or with one sequence per line are supported by default. Mismatch at the 3′ end is optional and the template DNA is set for human genome assembly GRCh37/hg19 by default. All other parameters have default values for routine analysis.

Running time

The running time is related to the number of input primer sequences. In general, the running time is more for large number of primer sequences. It usually takes <3 s for two primers, even with a large genome (e.g. human genome) as the template DNA. Several minutes are required for processing multiple primer sequences. It is to be noted that a maximum of 50 primer sequences are allowed by the online web server, whereas there is no such limitation for the command-line tool. A privacy link is created for later access when the task requires long running time.

Output

The MFEprimer-3.0 result page comprises six sections: (i) query: the list of user input primer sequences with information on binding site [15], primer size, and GC content, and T_m value annotation; (ii) hairpin list with detailed alignment structures; (iii) dimer list with detailed alignment structures; (iv) brief description of all the potential amplicons. Several buttons are also available for further analysis of amplicons, including 'Virtual Electrophoretogram', 'Sequence Alignment' and 'Output Amplicons'; (v) amplicon details: the detailed hybridization information of each predicted amplicon as well as the amplicon sequence. For the command-line tool, the SNPs are marked at the binding sites; (vi) parameters: the list of parameters used in the evaluation process. A protocol and a detailed tutorial can be found at https://www.mfeprimer.com/docs/.

THE MFEPRIMER-3.0 COMMAND-LINE TOOL

Compared to the web server, the MFEprimer-3.0 command-line tool has no background database- and primer sequence-related limitations. More importantly, MFEprimer-3.0 can output in the JSON format, which can be easily parsed by programs in PCR primer design pipelines. A primer design pipeline should contain a decision-making module (DMM) to allow it to accept or reject the candidate primers. The DMM is dependent on certain applications, such as multiplex PCR. For each application, users can create their own DMM and set appropriate cut-off values for specificity, dimers, and so on. For example, an accepted single-plex PCR primer pair should generate a specific target amplicon, but no dimers and hairpins, and should not have large number of stable binding sites [15,16]. Accepted multiplex PCR primers should not generate cross-dimers and non-specific amplicons [1]. Thus, in addition to the goal of PCR amplification, a primer design pipeline can be developed for automatically evaluating primer quality [17,18]. Figure 1
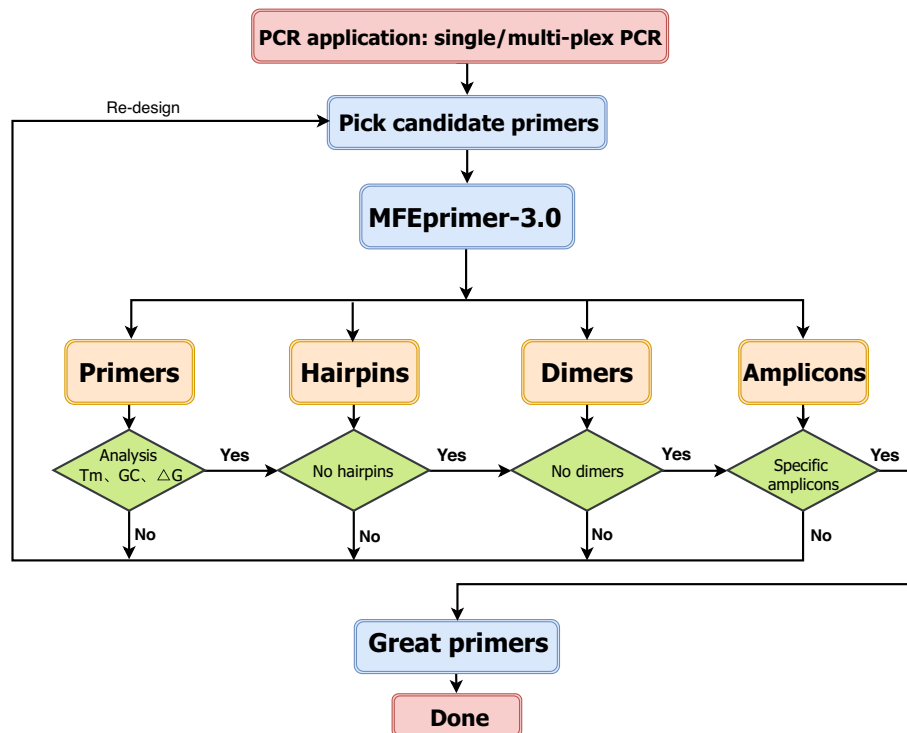
Figure 1. A single-plex/multiplex PCR primer design pipeline. MFEprimer-3.0 acts as a primer quality control module. First, several candidate primers are designed for the first template sequence using the primer pick module. Second, each pair of candidate primer is evaluated using MFEprimer-3.0 for dimers, hairpins, and specificity. Third, the primer pair is rejected if it fails to pass the DMM strategy. Otherwise, it is added to the 'great primers' pool. Fourth, for multiplex PCR, the cycle is repeated by returning to the first step for the second template sequence. Finally, a pool of primers is returned for all the template sequences.

shows a typical primer design pipeline for single-plex PCR and multiplex PCR. First, we used the primer design software, for example, Primer3 (11), to select several candidate primer pairs. Second, for each candidate pair of primers, we used MFEprimer-3.0 to analyze its specificity, dimers, and hairpins. Cross-dimers and non-specific amplicons must be evaluated for multiplex PCR. Third, for single-plex PCR, a primer pair is accepted if it generates the target amplicon but does not generate any other non-specific amplicons. For multiplex PCR, the candidate pair of primers is accepted if it does not generate any cross-dimers and non-specific amplicons. In addition, the number and stability of the binding sites for each primer were also important predictors of a failed PCR (15). Fourth, this process has to be repeated for multiplex PCR primer design until all the templates are designed successfully.

CONCLUSION

Multiplex PCR is an efficient DNA capture method for simultaneously amplifying up to thousands of SNPs in a single reaction, and multiplex PCR-based next-generation sequencing is widely used in biological and medical applications. This process requires compatible primer sets without non-specific amplifications and dimers. As it is not possible to manually analyze the specificity of two primers in one reaction, it is essential to develop a primer design pipeline that can be used to automatically evaluate primer quality. MFEprimer-3.0 is a functional primer QC program

for checking non-specific amplicons, dimers, hairpins, and other structures. The command-line tool with JSON output renders it more useful, as it can be integrated into most primer design pipelines as a QC module. In the future, we will continue to update the MFEprimer to make it a more powerful and effective PCR primer QC program.

REFERENCES

1. Shen,Z., Qu,W., Wang,W., Lu,Y., Wu,Y., Li,Z., Hang,X., Wang,X., Zhao,D. and Zhang,C. (2010) MPprimer: a program for reliable multiplex PCR primer design. *BMC Bioinformatics*, **11**, 143.
2. Park,N. and Vassiliou,G. (2017) Design and application of multiplex PCR Seq for the detection of somatic mutations associated with myeloid malignancies. *Methods Mol. Biol.*, **1633**, 87–99.
3. Ganschow,S., Silvery,J. and Tiemann,C. (2018) Development of a multiplex forensic identity panel for massively parallel sequencing

and its systematic optimization using design of experiments. *Forensic Sci. Int. Genet.*, **39**, 32–43.

4. Sambo,F., Finotello,F., Lavezzo,E., Baruzzo,G., Masi,G., Peta,E., Falda,M., Toppo,S., Barzon,L. and Di Camillo,B. (2018) Optimizing PCR primers targeting the bacterial 16S ribosomal RNA gene. *BMC Bioinformatics*, **19**, 343.

5. Rosenfeld,R., Zvi,A., Winter,E., Hope,R., Israeli,O., Mazor,O. and Yaari,G. (2019) A primer set for comprehensive amplification of V-genes from rhesus macaque origin based on repertoire sequencing. *J. Immunol. Methods*, **465**, 67–71.

6. Andreson,R., Reppo,E., Kaplinski,L. and Remm,M. (2006) GENOMEMASKER package for designing unique genomic PCR primers. *BMC Bioinformatics*, **7**, 172.

7. Ye,J., Coulouris,G., Zaretskaya,I., Cutcutache,I., Rozen,S. and Madden,T.L. (2012) Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. *BMC Bioinformatics*, **13**, 134.

8. Qu,W., Shen,Z., Zhao,D., Yang,Y. and Zhang,C. (2009) MFEprimer: multiple factor evaluation of the specificity of PCR primers. *Bioinformatics*, **25**, 276–278.

9. Qu,W., Zhou,Y., Zhang,Y., Lu,Y., Wang,X., Zhao,D., Yang,Y. and Zhang,C. (2012) MFEprimer-2.0: a fast thermodynamics-based program for checking PCR primer specificity. *Nucleic Acids Res.*, **40**, W205–W208.

10. Vallone,P.M. and Butler,J.M. (2004) AutoDimer: a screening tool for primer-dimer and hairpin structures. *BioTechniques*, **37**, 226–231.

11. Untergasser,A., Cutcutache,I., Koressaar,T., Ye,J., Faircloth,B.C., Remm,M. and Rozen,S.G. (2012) Primer3—new capabilities and interfaces. *Nucleic Acids Res.*, **40**, e115.

12. Koressaar,T., Lepamets,M., Kaplinski,L., Raime,K., Andreson,R. and Remm,M. (2018) Primer3_masker: integrating masking of template sequence with primer design software. *Bioinformatics*, **34**, 1937–1938.

13. Ye,S., Dhillon,S., Ke,X., Collins,A.R. and Day,I.N. (2001) An efficient procedure for genotyping single nucleotide polymorphisms. *Nucleic Acids Res.*, **29**, E88.

14. SantaLucia,J. and Hicks,D. (2004) The thermodynamics of DNA structural motifs. *Annu. Rev. Biophys. Biomol. Struct.*, **33**, 415–440.

15. Andreson,R., Mols,T. and Remm,M. (2008) Predicting failure rate of PCR in large genomes. *Nucleic Acids Res.*, **36**, e66.

16. Qu,W. and Zhang,C. (2015) Selecting specific PCR primers with MFEprimer. *Methods Mol. Biol.*, **1275**, 201–213.

17. Xie,S., Zhu,Q., Qu,W., Xu,Z., Liu,X., Li,X., Li,S., Ma,W., Miao,Y., Zhang,L. *et al.* (2018) sRNAPrimerDB: comprehensive primer design and search web service for small non-coding RNAs. *Bioinformatics*, **35**, 1566–1572.

18. Youngblut,N.D., Barnett,S.E. and Buckley,D.H. (2018) SIPSim: a modeling toolkit to predict accuracy and aid design of DNA-SIP experiments. *Front. Microbiol.*, **9**, 570.