Bioinformatics, 34(11), 2018, 1937–1938
doi: 10.1093/bioinformatics/bty036
Advance Access Publication Date: 19 January 2018
Applications Note

Sequence analysis

Primer3_masker: integrating masking of template sequence with primer design software

Triinu Kõressaar, Maarja Lepamets, Lauris Kaplinski, Kairi Raime, Reidar Andreson and Maido Remm*

Department of Bioinformatics, Institute of Molecular and Cell Biology, University of Tartu, Tartu, Estonia

*To whom correspondence should be addressed.
Associate Editor: John Hancock

Received on October 16, 2017; revised on January 8, 2018; editorial decision on January 17, 2018; accepted on January 18, 2018

Abstract

Summary: Designing PCR primers for amplifying regions of eukaryotic genomes is a complicated task because the genomes contain a large number of repeat sequences and other regions unsuitable for amplification by PCR. We have developed a novel *k*-mer based masking method that uses a statistical model to detect and mask failure-prone regions on the DNA template prior to primer design. We implemented the software as a standalone software *primer3_masker* and integrated it into the primer design program Primer3.

Availability and implementation: The standalone version of *primer3_masker* is implemented in C. The source code is freely available at https://github.com/bioinfo-ut/primer3_masker/ (standalone version for Linux and macOS) and at https://github.com/primer3-org/primer3/ (integrated version). Primer3 web application that allows masking sequences of 196 animal and plant genomes is available at http://primer3.ut.ee/.

Contact: maido.remm@ut.ee

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Second generation sequencing technologies have tremendously increased the list of sequenced genomes. Consequently, specific PCR primers are frequently needed for amplifying certain regions from a wide range of organisms. Increasingly popular is sequencing of various non-model organisms such as marine invertebrates and various plants (Cannarozzi *et al.*, 2014; Picq *et al.*, 2014; Sanseverino *et al.*, 2015; Suresh *et al.*, 2014). Most of these organisms have large genomes with a high fraction of repetitive sequences. When amplifying regions from large eukaryotic genomes, the PCR primer failure rate is a notable problem. In our previous work, we have shown that the best predictor for PCR failure is the number of potential primer binding sites (Andreson *et al.*, 2008). To address the problem of non-specific binding sites, the primer design for large genomes typically requires executing independent software (Fig. 1, paths A and B) that either masks the unsuitable genomic regions prior to primer design (Andreson *et al.*, 2006a,b; Bedell *et al.*, 2000; Morgulis *et al.*, 2006) or runs a homology search after the design process to exclude primers with non-specific binding sites (Ye *et al.*, 2012). However, running two or more different programs is inconvenient and often does not justify itself. Also, generic homology search software is not optimized for searching primer binding sites from large genomic sequences and do not account for the asymmetric properties of primer sequences (matches to primer 3′-end are scored similarly to primer 5′-end despite their different binding properties on template DNA).

Other possible approaches for avoiding primers with non-specific binding sites are the use of species-specific mispriming libraries, which are integrated into Primer3 for a handful of model organisms (human, rodent and *Drosophila*; Fig. 1, path C), or the use of a built-in search for mispriming on template DNA while inputting the whole genome as the template. Neither of those choices is optimal since the former takes account only the repeats that have previously been discovered and the latter is very inefficient and time-consuming, particularly if the thermodynamic modelling of binding sites is enforced. Thus, a more efficient solution is required.

We have created a *k*-mer based software *primer3_masker* that masks the regions of template DNA that correspond to the binding sites

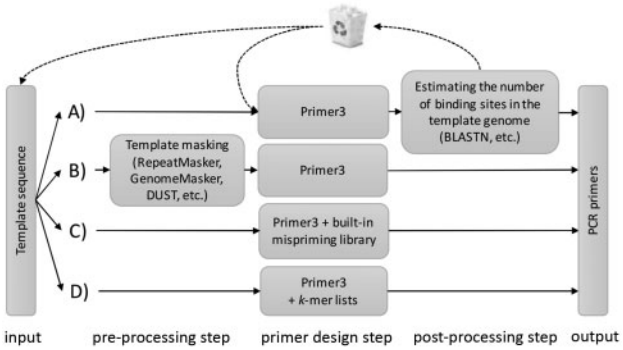

Fig. 1. Alternative methods for primer design and validation. Previous solutions for designing genomic primers require multiple steps (paths A–B) or species-specific mispriming library (path C). Primer3 with the novel masking function (path D) allows designing of PCR primers in only one step (for species that are covered with pre-made lists)

of the primers likely to fail in PCR reaction and integrated this functionality into Primer3 source code. We also provide pre-generated k-mer lists for 196 fully sequenced animal and plant genomes. We can provide help with generating lists from user-provided genome sequences.

2 Implementation

In our current work, we simplified our previously developed statistical model for prediction of PCR primer failure rate (Andreson *et al.*, 2008) by including only the frequencies of *k*-mers in a given genome as predictors. We implemented it in a software that can mask regions in DNA sequences based on their predicted failure rate in PCR. Using *k*-mers for modelling PCR primer binding sites provides a good compromise between speed, model complexity and model accuracy. The predictive value of the model was tested on the same experimental dataset and the same methodology as used previously (Andreson *et al.*, 2008) and found to perform similarly to previously described models (Supplementary Fig. S1).

The probability of PCR failure P_f for a given position in a genome sequence is calculated with the formula

$$P_f = \frac{e^{-4.336 + 0.1772 \times K11 + 0.239 \times K16}}{1 + e^{-4.336 + 0.1772 \times K11 + 0.239 \times K16}}$$

where K16 and K11 are natural logarithms of the respective genome-wide frequencies of the 16-mer and the 11-mer overlapping with the given position in their 3′-end (see Supplementary Fig. S2). K16 and K11 are retrieved from pre-built sorted *k*-mer frequency lists. Building of *k*-mer lists for newly sequenced genomes can be easily automated using the previously published *GenomeTester4* software package (Kaplinski *et al.*, 2015). We have prepared *k*-mer lists for 196 genomes.

For masking, the sliding window with step 1 and length 16 is used. The *k*-mer frequencies K16 and K11 are retrieved from the lists and P_f is calculated for each window. If P_f exceeds the cut-off (default = 0.10), then by default the 3′-end nucleotide in a given window is masked (Supplementary Fig. S3). All parameters including the full failure rate model can be adjusted. The standalone masker can mask up to 200 000 nucleotides per second on mid-range server hardware.

The same algorithm was integrated into the source code of a popular primer design software Primer3 (Untergasser *et al.*, 2012), allowing template masking and primer designing as a single-step process. Primer3 already has the functionality that allows using soft-masked template in the primer design (Koressaar and Remm, 2007). We added the functionality of masking the template DNA, as well as ordering the primer pairs in the final output according to their P_f values. All other design steps remained unchanged. In the integration process, we introduced six new global parameters to Primer3 (Supplementary Table S1). The masking feature is available from Primer3 core version 2.4 and web version 4.1.

3 The extent of masking in different genomes

The fraction of nucleotides that are masked is dependent on the organism and the cut-off value of P_f. At the default value $P_f = 0.10$, 44% of human genome is masked, whereas plant genomes are masked even more excessively: 76% of nucleotides in wheat and 71% in maize genomes are masked (Supplementary Table S2). The effect of P_f on masking of different genomes is shown in Supplementary Table S3 and Supplementary Figure S4. Standalone *primer3_masker* allows masking more than one nucleotide per window. This option can increase the reliability of designed primers at the expense of decreasing the usable fraction of the genome. Masking the entire window of 16 nucleotides increases the fraction of masked nucleotides typically by 10–20% (Supplementary Table S4). Examples of the masking extent and style of *primer3_masker* and RepeatMasker are shown in Supplementary Table S2 and Supplementary Figure S5.

4 Conclusions

The main strengths of *primer3_masker* are: (i) repeat detection is based on a statistical model, designed specifically for the identification of regions that might be problematic in PCR; (ii) repeat detection is based on *k*-mer frequency and can be easily applied for any sequenced genome; (iii) integration with web-based primer design software.

Funding

This work was funded by institutional grant IUT34-11 from the Estonian Ministry of Education and Research and the EU ERDF grant No. 2014-2020.4.01.15-0012 (Estonian Center of Excellence in Genomics and Translational Medicine).

Conflict of Interest: none declared.

References

Andreson,R. *et al.* (2006a) GENOMEMASKER package for designing unique genomic PCR primers. *BMC Bioinformatics*, **7**, 172.

Andreson,R. *et al.* (2006b) SNPmasker: automatic masking of SNPs and repeats across eukaryotic genomes. *Nucleic Acids Res.*, **34**, W651–W655.

Andreson,R. *et al.* (2008) Predicting failure rate of PCR in large genomes. *Nucleic Acids Res.*, **36**, e66.

Bedell,J.A. *et al.* (2000) MaskerAid: a performance enhancement to RepeatMasker. *Bioinformatics*, **16**, 1040–1041.

Cannarozzi,G. *et al.* (2014) Genome and transcriptome sequencing identifies breeding targets in the orphan crop tef (*Eragrostis tef*). *BMC Genomics*, **15**, 581.

Kaplinski,L. *et al.* (2015) GenomeTester4: a toolkit for performing basic set operations – union, intersection and complement on *k*-mer lists. *Gigascience*, **4**, 58.

Koressaar,T. and Remm,M. (2007) Enhancements and modifications of primer design program Primer3. *Bioinformatics*, **23**, 1289–1291.

Morgulis,A. *et al.* (2006) WindowMasker: window-based masker for sequenced genomes. *Bioinformatics*, **22**, 134–141.

Picq,S. *et al.* (2014) A small XY chromosomal region explains sex determination in wild dioecious V. vinifera and the reversal to hermaphroditism in domesticated grapevines. *BMC Plant Biol.*, **14**, 229.

Sanseverino,W. *et al.* (2015) Transposon insertions, structural variations, and SNPs contribute to the evolution of the melon genome. *Mol. Biol. Evol.*, **32**, 2760–2774.

Suresh,B.V. *et al.* (2014) Tomato genomic resources database: an integrated repository of useful tomato genomic information for basic and applied research. *PLoS One*, **9**, e86387.

Untergasser,A. *et al.* (2012) Primer3 - new capabilities and interfaces. *Nucleic Acids Res.*, **40**, e115.

Ye,J. *et al.* (2012) Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. *BMC Bioinformatics*, **13**, 134.