BIOINFORMATICS APPLICATIONS NOTE
Vol. 27 no. 2 2011, pages 266–267
doi:10.1093/bioinformatics/btq624

Genome analysis
Advance Access publication November 9, 2010

PriSM: a primer selection and matching tool for amplification and sequencing of viral genomes

Qing Yu[1], Elizabeth M. Ryan[1], Todd M. Allen[2], Bruce W. Birren[1], Matthew R. Henn[1] and Niall J. Lennon[1,*]

[1]Broad Institute of MIT & Harvard, Cambridge, MA 02142 and [2]Ragon Institute of MGH, MIT & Harvard,
149 13th Street Charlestown, MA 02129, USA

Associate Editor: Alfonso Valencia

ABSTRACT

Summary: PriSM is a set of algorithms designed to select and match degenerate primer pairs for the amplification of viral genomes. The design of panels of hundreds of primer pairs takes just hours using this program, compared with days using a manual approach. PriSM allows for rapid *in silico* optimization of primers for downstream applications such as sequencing. As a validation, PriSM was used to create an amplification primer panel for human immunodeficiency virus (HIV) Clade B.

Availability: The program is freely available for use at: www.broadinstitute.org/perl/seq/specialprojects/primerDesign.cgi

Contact: nlennon@broadinstitute.org

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

Received on July 26, 2010; revised on October 11, 2010; accepted on November 3, 2010

1 INTRODUCTION

Viral populations are dynamic and the strains circulating in an infected population often vary geographically and temporally. RNA viruses, especially those that cause chronic disease such as hepatitis C virus (HCV) and human immunodeficiency virus (HIV), exhibit tremendous sequence diversity (Brumme *et al.*, 2008; Martell *et al.*, 1992). As such, primers designed from a single reference genome may limit the number of variants captured from within a single patient as well as variants isolated from different geographic regions and time periods. Here, we report on a novel primer design method that: (i) leverages existing genome sequence information to design degenerate primers that account for the significant diversity observed in viral populations, and (ii) maximizes the ability to successfully amplify and sequence the genomes of viruses from different individuals that may differ significantly from a single viral reference genome.

Previous primer selection algorithms have been designed to amplify single target regions (Rozen and Skaletsky, 2000), or optimized for SNP or gene-specific PCR (Tsai *et al.*, 2007), group-specific PCR (Jarman, 2004), or for phylogenetically related DNA sequences (Fredslund *et al.*, 2005). These methods do not, however, permit push button, full-genome length, primer pair design from very diverse organisms such as those represented by RNA viruses. A comparison of the features of PriSM to those of several other primer design programs is included in Supplementary Table 1.

PriSM takes advantage of existing sequence information for a given viral strain to select the optimal primers, based on user-defined criteria. The algorithm then suggests pairs of primers to make up a panel that will generate amplicons of desired length and overlap. Diversity capture is increased through the use of degenerate base codes.

2 METHODS

The user interface allows for customization of the primer and amplicon selection parameters. Once the parameters are designated, the user can choose to upload either: (i) a multiple alignment of full-length genome sequences for the viral species of interest, created using MUSCLE (www.ebi.ac.uk/Tools/muscle/) or CLUSTALW (www.ebi.ac.uk/Tools/clustalw2/), and the multiFASTA used to create the alignment, or (ii) a multiFASTA file containing all the genomes used to create the multiple alignment file. In the latter case, the alignment is automatically created using CLUSTALW, with default settings. Finally, the user inputs an email address to which the output files will be sent. The process through which PriSM selects and matches primers has three major steps: make consensus, primer design and amplicon selection (see Supplementary Material, Fig. 1, for a detailed workflow).

Make consensus: PriSM first identifies the start and end positions of the alignment in the multiple alignment file. It then calculates the frequency of each base at each position of the alignment. Additional information pertaining to the handling of the end regions of the alignment, positions with missing bases and the files that are generated at each step can be found in the Supplementary Material. *Primer design*: starting with the majority consensus, primer candidates are chosen through the iterative evaluation of subsequence windows. Each subsequence window is a primer candidate that then is assessed for validity (see Supplementary Material for the full list of criteria used to select valid primers including T_m calculation and self-complementarity assessment). *Amplicon selection*: valid primers exist in two sets—Forward primers and Reverse primers. Starting with the Forward primer that begins at the most 5′ position of the genome each member of the set of Reverse primers is interrogated to find a partner that has less than a user-defined amount of complimentary sequence (default is 15 bases) and produces a product that is larger than the minimum amplicon length (default 500 bases) and smaller than the maximum amplicon length (default 900 bases). Once successfully paired with a Forward primer, a given Reverse primer is excluded from further pairing. Primer pairs are examined to evaluate the overlap between adjacent amplicons. Pairs that overlap by less than the user-defined threshold are removed (default 100 bases). Finally, pairs are ranked according to the combined conservation score of both members. A description of the files sent to the user is included in Supplementary Material.

*To whom correspondence should be addressed.

3 RESULTS

We have used PriSM to design successful primers suitable for the full-genome sequencing of Dengue Virus (DENV), West Nile Virus (WNV), HIV, HCV and Lassa Virus (data not shown). All of these viruses exist natively as single-stranded RNA genomes with genome lengths in the 9–11 kb range. Here, we report on performance results from HIV Clade B as this demonstrates the stability of PriSM even when viral strain diversity is high.

We downloaded 86 full-genome sequences for HIV Clade B from the Los Alamos HIV Database (www.hiv.lanl.gov/) and a multiple alignment was created using the MUSCLE algorithm with standard settings. Using default settings of PriSM, 1051 Forward primers and 1047 Reverse primers were deemed to be valid. From these, 352 amplicons between 500 and 900 bp were identified. To meet our coverage target and fit on our plate-based PCR format, a total of 96 of these were chosen to amplify the HIV genome (See Supplementary Material for a list of the primer pairs). The computational part of the process took 2 h 45 min between job submission and receipt of an email with the final results.

In order to test the robustness of the PriSM-designed primers, HIV RNA was extracted from 10 chronically infected patients and converted to cDNA as previously described (Brumme *et al.*, 2008). A single large amplicon spanning nucleotides 623–9639 of the HIV genome was generated from each cDNA and nested PCR reactions were set up using the PriSM-designed panel. Eight out of ten samples had amplification success rates (as calculated by number of successful products out of 82 possible) of >90%, with another sample yielding >85% positive bands. Only one sample had three or more overlapping products fail (see Supplementary Material for more details). These results compare very favorably with our observed HIV PCR performance historically (data not shown).

4 CONCLUSIONS

PriSM is the first program that we are aware of that is specifically designed to create a panel of primer pairs that tile along the entire length of a viral genome. PriSM leverages *a posteriori* information from viral sequence databases to build amplification primer pairs in a fraction of the time it would take an individual to do so manually. PriSM is a highly flexible algorithm that requires only FASTA files or alignments as inputs, making it easy to incorporate new data as they become available.

Our experience using this program on HIV as well as several other viruses indicates that PriSM is a useful, timesaving tool for researchers who wish to amplify full-length viral genomes.

ACKNOWLEDGEMENTS

The authors would like to thank Brian Haas and Robert Lintner for manuscript review and Boris Bugalter, Tim Ledlie and Michael Koehrsen for early contributions to the project.

Funding: This project has been funded in part with Federal funds from the National Institute of Allergy and Infectious Disease, National Institutes of Health, Department of Health and Human Services, under Contract No. HHSN266200400001C (to B.W.B.), and through the Collaboration for AIDS Vaccine Discovery with partial support from the Bill & Melinda Gates Foundation.

Conflict of Interest: none declared.

REFERENCES

Brumme,Z.L. *et al*. (2008) Marked epitope- and allele-specific differences in rates of mutation in human immunodeficiency type 1 (HIV-1) Gag, Pol, and Nef cytotoxic T-lymphocyte epitopes in acute/early HIV-1 infection. *J. Virol.*, **82**, 9216–9227.

Fredslund,J. *et al.* (2005) PriFi: using a multiple alignment of related sequences to find primers for amplification of homologs. *Nucleic Acids Res.*, **33**, 20.

Jarman,S.N. (2004) Amplicon: software for designing PCR primers on aligned DNA sequences. *Bioinformatics*, **20**, 1644–1645.

Martell,M. *et al.* (1992) Hepatitis C virus (HCV) circulates as a population of different but closely related genomes: quasispecies nature of HCV genome distribution. *J. Virol.*, **66**, 3225–3229.

Rozen,S. and Skaletsky,H.J. (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.* Humana Press, Totowa, NJ, pp. 365–386.

Tsai,M. *et al.* (2007) PrimerZ: streamlined primer design for promoters, exons and human SNPs. *Nucleic Acids Res.*, **35**, W65.