DOI 10.18699/SBB-2022-195

Detecting broad specter of viral nucleic acids via NGS: novel k-mer-based algorithm for multiplex degenerate primer sets design

Matsvay A.D.*, Stetsenko I.F., Ayginin A.A., Mikhaylov I.M., Bogacheva A.I., Say A.V., Shipulin G.A.

Federal State Budgetary Institution "Centre for Strategic Planning and Management of Biomedical Health Risks" of the Federal Medical Biological Agency, Moscow, Russia

* *AMatsvay@cspmz.ru*

Key words: multiplex PCR, primer design, k-mer, algorithm

Motivation and Aim: Multiplex PCR is a widely used method of target enrichment prior to library preparation for next generation sequencing. Compared to other techniques, it is time-efficient, cheap and provides very high proportion of on-target reads. However, designing of primer panel for multiplex PCR is challenging. In case of applying multiplex PCR for detection of microorganism, a commonly used approach is constructing a multiple sequence alignment in order to identify sequence regions conserved between organisms of interest using, e.g., ClustalW [1]. However, if genomes of organisms of interest are diverged enough, which is a common case for taxonomically distant viruses, construction of MSA may not be possible at all, or the resulting alignment may be not suitable for primer design.

Methods and Algorithms: A newly developed algorithm avoids this issue and allows designing a primer panels for parallel detection of diverged viral taxonomic groups without MSA construction. It utilizes k-mer approach and automatically selects degenerate primers, considering their thermodynamic characteristics, representative database coverage, predicted sizes of amplified sequences, stability of homo- and hetero-dimeric structures and hairpins and the stability of 3' end annealing. Relevant degenerate positions are introduced during the early stages of the algorithm, which allows to calculate unbiased statistics at subsequent steps. Using GPU parallelization for matrix calculation allows processing large databases containing highly diverged genomes of phylogenetically distant groups of viruses, which was shown to be impossible with other existing k-mer-based primer design software [2].

Results: As a proof of concept, the algorithm was used to design a primer panel targeting all known viruses causing respiratory infection. As predicted by *in silico* modelling, panel provides 99.81 % coverage of all known representatives (16 genera of 10 different families). Compared to primer sets described earlier, the newly developed primer set provides a number of benefits, including smaller degenerate position count and broader viral genome coverage per single structure within primer pair. The primer was used to prepare NGS libraries form a number of control samples, containing human adenoviruses (types 1–7, 21), bocavirus, alpha- (229E) and betacoronaviruses (SARS-CoV-2, type 1), influenza B virus, measles virus, meta- and orthopneumovirus, para-influenza virus (types 1–3) and rhinovirus (types 13, 15, 17, 26, 29) nucleic acids, and the sequencing results were 100 % concordant with viral identification by real time PCR-based test systems.

References

1. Chenna R. et al. *Nucleic Acids Res.* 2003;31(13):3497-3500.
2. Hysom D.A. et al. *PloS One*. 2012;7(4):e34560.