Validation of novel hereditary cancer genes identified upon exome sequencing: a focus on primer design

I.V. Bizin
Department of Bioinformatics, Peter the Great St. Petersburg Polytechnic University, St. Petersburg, Russia, bizin@yandex.ru

A.P. Sokolenko, E.Sh. Kuligina
Department of Tumor Growth Biology, N.N. Petrov Institute of Oncology, St. Petersburg, Russia

Whole exome sequencing (WES) opened new opportunities for the identification of novel genes for hereditary human diseases. WES studies usually produce hundreds or thousands candidates, which require subsequent validation in cases-controls studies. The analysis of each candidate includes individual PCR design for performing of allele-specific polymerase chain reaction (PCR), high-resolution melting (HRM) and Sanger sequencing [1]. For the time being, the selection of primers for PCR remains a time consuming procedure. For example, none of the available software instruments (Primer-BLAST, Primer3, ThermonucleotideBLAST [2,3,4]) allows to generate a pool of primers for the entire coding region of the gene, therefore each coding fragment has to be selected and analyzed manually. The existing tools have limited capability to ensure, the primers will not amplify a non-target fragment of genome, therefore the check for their specificity involves manual labor. The consideration of single nucleotide polymorphisms lying within primer sequence is usually performed without adjustment for their frequency. We aimed to develop a pipeline, which would facilitate the process of large-scale primer design in automatic mode. The pipeline combines the advantageous feature of available software tools.

In particular, we created automatic upload of the sequence of interest with flanking regions using protein_id or genome coordinates from reference genome. For this step we used Biopython tools [5].

We utilized the fragment of the ThermonucleotideBLAST instrument in order to ensure PCR specificity. While other computer tools focus main on a heuristic definition of sequence similarity, this software runs in-silico PCR based on thermodynamic similarity and minimizes the risk of the amplification from a non-target template.

Our software detects potential polymorphic sites using public databases (dbSNP, Exome Aggregation Consortium, NHLBI Exome Sequencing Project). If the primer falls on the SNP-containing fragment, the frequency of this allelic variation is considered. Also, the position of the SNP within potential primer (5'-end or 3'-end) is analyzed. Overall, if the SNP is rare and located at the 5'-end of the primer, it is highly unlikely to compromise further experiments.

Our software tool is also capable to suggest the design for the analysis of long DNA sequences. It considers several variants of breaking the entire sequence for overlapping DNA fragments, and creates the sets of primers for each potential amplicon. Then it analyzes, which of the suggested fragments contains the most reliable primers, i.e. those which are unlikely to form dimers, recognize the non-target sequence, or be compromised by SNPs. As result, the design for the analysis of entire coding sequence of a gene is generated in an automatic way.

The utility of this tool was validated in a whole exome sequencing study for hereditary breast cancer, which utilized 173 candidate mutations and entire coding regions of 12 candidate genes.

The research has been supported by Russian Science Foundation, project 16-45-02011.

1. A.P.Sokolenko et al. (2015) Identification of novel hereditary cancer genes by whole exome sequencing, Cancer Letters, **369(2):**274–288.

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

3. J.D.Gans and M.Wolinsky (2008) Improved assay-dependent searching of nucleic acid sequence databases, Nucleic Acids Res., **36(12):**e74

4. A.Untergasser et al. (2012) Primer3 – new capabilities and interfaces, Nucleic Acids Res., **40(15):**e115.

5. P.A.Cock et al. (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics, Bioinformatics, **25:**1422-1423