ConsensusPrime - A Bioinformatic Pipeline For Ideal Consensus Primer Design

Maximilian Collatz (maximilian.collatz@leibniz-ipht.de)
Leibniz Institute of Photonic Technology (IPHT), 07743 Jena, Germany

Sascha D. Braun
Leibniz Institute of Photonic Technology (IPHT), 07743 Jena, Germany

Stefan Monecke
Leibniz Institute of Photonic Technology (IPHT), 07743 Jena, Germany

Ralf Ehricht
Leibniz Institute of Photonic Technology (IPHT), 07743 Jena, Germany

Research Article

Keywords: ConsensusPrime, molecular, sequence, homogeneous

Posted Date: November 2nd, 2021

DOI: https://doi.org/10.21203/rs.3.rs-1030641/v1

Abstract

High quality oligonucleotides for molecular amplification and detection procedures of diverse target sequences depend on sequence homology. The identification of homogeneous regions in alignments can be done manually only if alignments are small and if they contain sequences of high similarity. For large and inhomogeneous alignments, the process of finding the best regions needs to be automated.

The ConsensusPrime pipeline was developed to detect the most homolog regions from multiple sequences. It automates the prediction of optimal consensus primers for analytical and diagnostic procedures/assays.

ConsensusPrime is a fast and easy to use pipeline for predicting optimal consensus primers that is executable on local systems without depending on external resources and webservices. An implementation in a Docker image ensures platform-independent executability and installability despite the combination of multiple programs. The source code and installation instructions are freely available on GitHub (https://github.com/mcollatz/ConsensusPrime).

Background

The basic unit of organization of all living organisms is the cell, and every cell carries a genome which encodes all of its properties in their DNA. In the recent 20 years, it has become feasible to read these genomes as a whole. Comparison of complete genomes is now becoming the gold standard for genotyping, and knowledge of full-length genome sequences is becoming the basis for developing various molecular assays for detection and typing. DNA is a long-term stable molecule and can be amplified exponentially (PCR, isothermal amplifications) enabling highly sensitive assay formats. The specific base pair binding in DNA – DNA and DNA – RNA hybridizations allow a highly specific detection of nucleic acid sequences even in low concentration ranges and under stringent conditions. Therefore, molecular assays can easily be developed, manufactured, stored, delivered, multiplexed and compared within different labs with a reasonable price/test. But there are also challenges in the assay development. The design of oligonucleotide primers and/or probes is a crucial task for many of these above-mentioned biomedical studies and molecular applications aiming, e.g., on the identification of pathogenic organisms and their resistance and virulence genes. Other applications benefitting from high-quality consensus oligonucleotides include FISH analysis, Northern blots, microarray analyses, isothermal amplification procedures, and all types of molecular assays utilizing linear or exponential amplifications. To ensure the widespread usability of oligonucleotide sequences, the detection of high-quality consensus regions among relevant target sequences is a necessary pre-requisite. Given the inherent variability of, e.g., bacterial target genes and given the massive increase of published target sequences, it becomes necessary to automatize the process of aligning these sequences and detecting suitable consensus regions.

For the performance and applicability of the predicted primers, the selection of the target sequences that serve as the basis for the alignment plays a decisive role. This is because functional oligonucleotide sequences can only be predicted if the selected sequences are also a representative image of the sequences to be examined in a given sample type. The basis for a representative consensus sequence is a high-quality multiple sequence alignment. The alignment quality depends on several factors such as type and length of the input sequences, parameters like gap opening/extension costs and on the algorithm itself. Therefore, the parameters for calculating the alignments vary based on the particular requirements of the alignment.

Materials And Methods

All processing steps, including the input and output of files and parameters, are visualized as a flowchart in Figure 1. As the entire pipeline is based on multiple sequence alignments, their quality is of great importance. Therefore, the parameters of MAFFT [1] are adjusted to precisely fulfill the alignment requirements in every alignment step. This is of particular importance when aligning the short primer sequences for visualization. In this case the - - adda$gments$ parameter of MAFFT is used to properly align the short primers to their origin. MAFFT also allows the automated adjustment of the strand direction of a sequence. Another important parameter is - - $adjustdirectio \in$ that allows the automated detection and adjustment of the strand direction in which sequences are provided, as well as the mapping of the reverse primers.

To avoid unwanted distortions of the consensus score(s) due to overrepresented sequences, identical and partial sequences are first removed from the alignment. The removal can be influenced via the - - $gapthreshold$ parameter by providing a value between 0 and 1. The default value of 0.2 results in the removal of all sequences of the alignment that have more than 20% gap symbols. To identify ideal consensus oligos, the consensus sequence is needed. Therefore, the pipeline uses MAFFT to align all input sequences together in a global multiple sequence alignment and it identifies for every position in the alignment the most common nucleotide. In addition, a consensus score is calculated for every alignment position which is the ratio of the respective count/number of most common nucleotide or gap symbol (-) at that position to the total number of sequences. All letters that are not ATGC are treated as gap. A perfectly conserved region in which all sequences at a given position are identical is thus assigned a consensus score of 1. The pipeline allows the user to control the quality values of the consensus sequence used for primer prediction via the - - $consensusthreshold$ parameter. The default value of 0.95 ensures that the most abundant nucleotide occurs in at least 95% of the sequences at the given position. In addition, the regions above the threshold must have a contiguous minimum length of at least 20 nucleotides. All regions that fall below these values are excluded from the subsequent primer prediction.

Before the consensus regions are identified for primer design, any gaps are removed from the consensus sequence as well as the corresponding value from the consensus scores. This is necessary because gaps are not encoded by nucleotides and are therefore not relevant for primer design. Gaps in the consensus sequence are caused by insertions in one or more sequences. From this “gapless consensus sequence” the regions relevant for primer design are identified.

As Primer3 [2] searches for the primer pair in a contiguous sequence section, instead of using the area in which primers are to be searched (SEQUENCE_INCLUDED_REGION and SEQUENCE_INTERNAL_INCLUDED_REGION), all areas in which primers are not to be searched are excluded by the pipeline (SEQUENCE_EXCLUDED_REGION and SEQUENCE_INTERNAL_EXCLUDED_REGION). This allows the prediction of primers in non-consecutive sequence segments. Furthermore, the gapless consensus sequence is automatically written into the primer3 parameter file (SEQUENCE_TEMPLATE). All other parameters such as melting temperatures or primer lengths are taken from the user-defined Primer3 parameter file. (For a detailed parameter description check out the primer3 manual (https://primer3.org/manual.html). From this, Primer3 predicts the optimal consensus primers and displays the results in a plain text file. The ConsensusPrime pipeline reads the Primer3 output and creates a comprehensive output including all details of previous filter steps in .html format. The predicted primers are added to a final alignment to be visualized using ClustalX [3].

Results & Discussion

The ConsensusPrime pipeline allows automatic detection of optimal consensus regions from large alignments with many sequences. This is of particular interest for applications such as the prediction of real-time PCR primer and probes within regions of high homology. Adjustable physiochemical parameters can be set for the design of hundreds of sequences for DNA-based microarray analyses or sequences for the design of other molecular-based analyses (e.g., FISH, molecular beacon technology or isothermal amplifications). All parameters are defined in a simple text file, and therefore, the run parameters can be easily reused or adapted. All adjustments and parameters are summarized in the HTML overview for complete reproducibility.

The pipeline starts by aligning the input sequences in a multiple sequence alignment and the regions with the best homology are identified. These regions are then used for primer prediction. Subsequently, the predicted primer and probe sequences are visualized in a multiple sequence alignment.

Our new pipeline combines MAFFT (v7.453) [1] and Primer3 (v2.5.0) [2], two common and long-established tools to ensure high-quality alignments and, respectively, primer predictions.

The source code of the pipeline is written in python and easily executable from the Linux command line. The pipeline is also integrated in a ready-to-use Docker container which delivers every dependency pre-installed (mcollatz/consensusprime:1.0).

Embedding this pipeline in a Docker container ensures executability on a wide variety of systems. With regard to an independence from online servers, data security and patient privacy concerns are important arguments for a local executability, since the sequences do not have to be passed on to third parties.

Declarations

Authors' contributions

MC implemented the pipeline and wrote the manuscript. SB, SM and RE helped with conceptual development and requirements profiling.

All authors revised and corrected the various drafts and approved the final version of the manuscript.

Data availability

Project name: ConsensusPrime
Project home page: https://github.com/mcollatz/ConsensusPrime
Operating system(s): Linux (or independent with Docker)
Programming language: Python3.8
Other requirements: Mafft v7.453, primer3 v2.5.0, ClustalX (optional)
License: MIT license
Any restrictions to use by non-academics: no restrictions

Competing interests

The authors declare that they have no competing interests.

Funding

ADA (13GW0456C: BMBF): Adaptable decentralized diagnostics for veterinary and human medicine.

References

1 Nakamura, T., Yamada, K. D., Tomii, K. & Katoh, K. Parallelization of MAFFT for large-scale multiple sequence alignments. *Bioinformatics* **34**, 2490-2492, doi:10.1093/bioinformatics/bty121 (2018).

2 Untergasser, A. *et al.* Primer3–new capabilities and interfaces. *Nucleic Acids Res* **40**, e115, doi:10.1093/nar/gks596 (2012).

3 Larkin, M. A. *et al.* Clustal W and Clustal X version 2.0. *Bioinformatics* **23**, 2947-2948, doi:10.1093/bioinformatics/btm404 (2007).

Figures

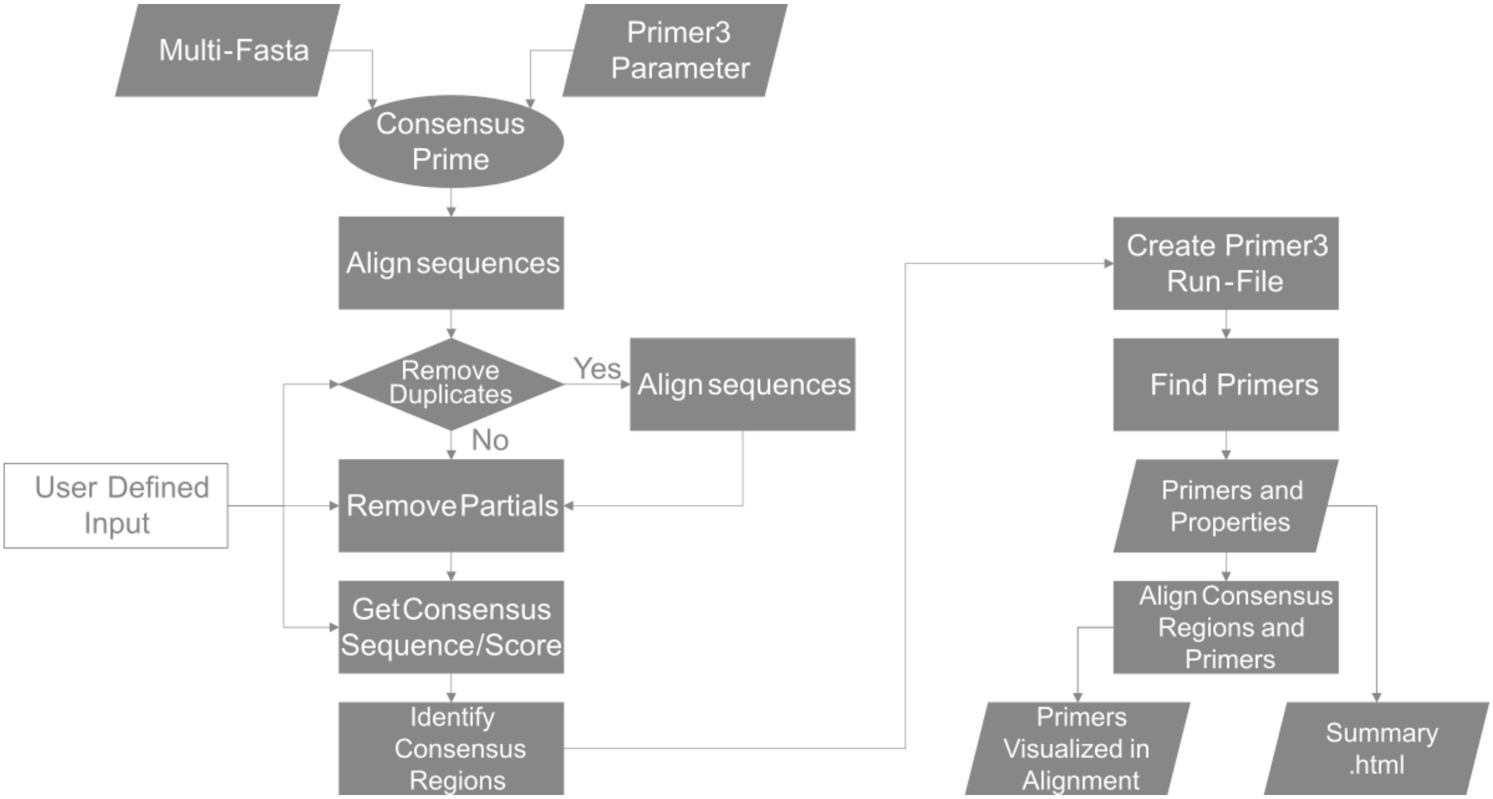

Figure 1

ConsensusPrime workflow with input/output files as well as user parameters. The sequences have to be provided in multi-fasta format. The Primer3 parameter file contains the primer/probe parameters for Primer3 and the ConsensusPrime parameters that influence the pipeline itself are given via the command line. The ConsensusPrime pipeline processes the input data in several successive filter and alignment steps to identify suitable consensus regions. The regions found in this way are automatically written to the Primer3 parameter file and the primer prediction is started using Primer3. Afterwards, the output of Primer3 in addition to the details of the individual filter steps are written to a concise .html file. Predicted primers are also added to a final alignment to be displayed with any alignment visualization tool.