Technical Note
ConsensusPrime—A Bioinformatic Pipeline for Ideal Consensus Primer Design

Maximilian Collatz 1,2,*, Sascha D. Braun 1,2, Stefan Monecke 1,2,3 and Ralf Ehricht 1,2,4

1 Leibniz Institute of Photonic Technology (IPHT), 07745 Jena, Germany
2 InfectoGnostics Research Campus, 07743 Jena, Germany
3 Institute for Medical Microbiology and Virology, Dresden University Hospital, 01307 Dresden, Germany
4 Institute of Physical Chemistry, Friedrich-Schiller University, 07743 Jena, Germany
* Correspondence: maximilian.collatz@leibniz-ipht.de

Abstract: Background: High-quality oligonucleotides for molecular amplification and detection procedures of diverse target sequences depend on sequence homology. Processing input sequences and identifying homogeneous regions in alignments can be carried out by hand only if they are small and contain sequences of high similarity. Finding the best regions for large and inhomogeneous alignments needs to be automated. Results: The ConsensusPrime pipeline was developed to sort out redundant and technical interfering data in multiple sequence alignments and detect the most homologous regions from multiple sequences. It automates the prediction of optimal consensus primers for molecular analytical and sequence-based procedures/assays. Conclusion: 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 web services. 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 publicly available on GitHub.

Keywords: consensus primer; primer design; alignment filtering; oligo design; molecular assays; bioinformatics

1. Background

The basic unit of organization of all living organisms is the cell; most cells carry a genome that encodes all its properties in their DNA. DNA (Deoxyribonucleic acid) is a polymer of two intertwined polynucleotide chains that form a double helix [1]. Information about an organism’s DNA allows detailed conclusions about its phenotypic characteristics. In the last 20 years, it has become feasible to sequence entire genomes rather than short fragments. Current third-generation sequencing methods, such as nanopore sequencing, have led to surging numbers of available genome sequences [2]. It is essential to incorporate this information into the design of future studies. Especially in the world of ever-changing pathogens, all available sequences should be integrated into the design of detection oligos or primers to ensure broad applicability.

Comparison and analysis 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. Compared to other biologically relevant macromolecules, DNA is a long-term stable molecule and can be amplified exponentially (PCR, isothermal amplifications), enabling highly sensitive assay formats [3]. The specific Watson–Crick base pair binding [4] in DNA—DNA and DNA—RNA hybridizations allows precise detection of nucleic acid sequences even in very low concentration ranges and under stringent conditions. Therefore, molecular assays can quickly be developed, manufactured, stored, delivered, multiplexed, and compared between different labs with a reasonable and constantly dropping price per test. However, there are also challenges in assay development. The design of oligonucleotide primers and...
probes is a crucial task for many biomedical studies and molecular applications aiming, e.g., to identify pathogenic organisms and their resistance and virulence genes. Other applications benefiting from high-quality consensus oligonucleotides include FISH analyses, Northern blots, microarray-based analyses, isothermal amplification procedures, and all types of molecular assays utilizing linear or exponential amplifications [5]. To ensure the widespread usability of oligonucleotide sequences, detecting high-quality consensus regions among relevant target sequences is a necessary pre-requisite. Given the inherent variability of, e.g., bacterial or viral target genes and the massive increase in the number 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 predicted primers, selecting 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 representative 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 the type and length of the input sequences, parameters such as gap opening/extension costs, and the algorithm itself. Therefore, the parameters for calculating the alignments vary based on the particular alignment requirements.

Consensus primer design is not a new problem. There are already publications on programs for the design of consensus primers. Unfortunately, these tools and their online services are often no longer available, such as CODEHOP [6]. Another tool for consensus primer design is PrimerDesign-M [7]. Unfortunately, this tool does not allow extensive parameter settings for experimental design, such as setting specific Tm values. A significant advantage of the ConsensusPrime pipeline is the various possibilities to filter the input alignment and thus improve the quality of the predicted primers.

2. Materials and Methods

With MAFFT [8] for generating multiple sequence alignments and Primer3 [9] for predicting primers and probes, the ConsensusPrime pipeline uses well-established tools. The pipeline needs three different pieces of input information. First, the sequences have to be provided in multi-fasta format. Second is the Primer3 parameter file containing the physicochemical primer/probe parameters for Primer3, and thirdly 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 into the Primer3 parameter file, and the primer prediction is started using Primer3. Afterward, the output of Primer3, in addition to the details of the individual filter steps, is written into a concise HTML file. Predicted primers are also added to a final alignment to be displayed with any alignment visualization tool.

The entire pipeline is based on multiple sequence alignments, so their quality is vital for functional assays. Therefore, the parameters of MAFFT were adjusted to 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 ‘–addfragments’ parameter of MAFFT is used to align the short primers to their origin properly. MAFFT also allows the automated adjustment of the strand direction of a sequence. Another important parameter is ‘–adjustdirection’, which 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 scores caused by overrepresented sequences, identical sequences are removed from the alignment unless specified differently using the ‘–keepduplicates’ parameter. The ‘–consensusimilarity’ parameter defines the similarity cutoff for each sequence in comparison to the consensus sequence of the alignment. The default value of, e.g., 0.8 means that a sequence has to have at least 80% of its
aligned nucleotides in common with the consensus sequence. Otherwise, the sequence is removed from the alignment for primer prediction. The filtered sequences are then re-aligned. Therefore, the pipeline uses MAFFT to align all input sequences in a global multiple-sequence alignment, and it identifies the most common nucleotide for every position in the alignment. 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 a 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 below these values are excluded from the subsequent primer prediction.

Before the consensus regions are identified for primer design, gaps are removed from the consensus sequence and the corresponding value from the consensus scores. This removal is necessary because gaps are not encoded by nucleotides and are irrelevant to primer design. Gaps in the consensus sequence are caused by insertions in one or more related sequences of the alignment for better visualization.

From this “gapless consensus sequence”, the regions relevant to the primer design are identified.

As Primer3 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, accessed on 7 October 2022). 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 [10] and can be manually inspected for logical errors. The structure of the final alignment is always in the order of filtered sequences, consensus sequence, sequence parts used for the primer prediction, and predicted primer/probe pairs. This alignment enables the user to check the alignment of the filtered input sequences and the resulting consensus regions considered for the actual primer prediction. It also gives an excellent overview of the predicted primers to choose the best pair if multiple predictions have been made. Note that the reverse primer is added to the final alignment as the reverse complementary sequence.

3. Results and Discussion

Unfortunately, automatically assembled sequence alignments often contain a large number of identical sequences as well as partial/fragmented sequences. The ConsensusPrime pipeline includes several automatic filtering and correction options to limit their bias. First, identical sequences are removed, and only a single copy is retained. Another important filtering option is to filter sequences based on their similarity to the consensus sequence. In other words, sequences that are very different from the majority of aligned sequences are removed. This includes sequences with a deviating nucleotide composition and partial sequences or sequences that contain large insertions compared to the consensus sequence of the input alignment. Short subsequences and very different sequences are often found in
automatically generated alignments in frequently used databases, which makes extensive filtering methods essential to ensure a reasonable data basis for the following analyses.

The ConsensusPrime pipeline allows the automatic detection of optimal consensus regions from large alignments with many sequences. The necessary alignment filter steps and functions are integral components of the pipeline and ensure an optimal input for the subsequent primer predictions. The created HTML overview provides a summary of the individual filter steps. In addition, all individual steps are stored as an alignment to ensure perfect traceability. Adjustable physiochemical parameters can be set for the design of hundreds of sequences for DNA-based microarray analyses or the design of other molecular-based analyses (e.g., FISH, molecular beacon technology, or isothermal amplification). All parameters are defined in a simple text file; therefore, the run parameters can be easily reused or adapted. All adjustments and parameters are summarized in the HTML overview for complete reproducibility. A final alignment is created for a clear visualization and inspection of the predicted primers. This alignment contains the input sequences used, the consensus sequence, the consensus regions representing the regions used for primer prediction, and all predicted primers. For an example, see Figure 1.

Figure 1. Final Alignment. Section of the final alignment visualized with ClustalX [10]. The vanA-A sequence of P. apiarius contains many nucleotides that deviate from the consensus sequence. In this case, the user must decide whether this sequence should remain in the alignment or remove it by adjusting the filter parameters.
A possible use case for the pipeline would be the creation of a PCR primer/probe set for the detection of an antibiotic resistance gene in a widespread bacterium with the aim of a microarray to detect the abundance of the gene [11]. Because the bacterium is widespread, the number of genotypic variations to be considered is correspondingly high. In order to design a universal primer/probe pair, the entirety of the available sequence information must be included in the design by using multiple sequence alignments. This is the only way to ensure that the primers/probes are located in relevant regions across all targeted sequences.

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.

Our new pipeline combines custom alignment filters with MAFFT (v7.453) [8] and Primer3 (v2.5.0) [9], two familiar and long-established tools to ensure high-quality alignments and, respectively, primer predictions. An overview of the pipeline is given as a flow diagram in Figure 2.

![Figure 2. ConsensusPrime workflow. Overview of the ConsensusPrime pipeline with all input/output files and user-provided parameters.](image)

The source code of the pipeline is written in python and easily executable from the Linux command line. To keep the pipeline as independent as possible from further dependencies, it was decided not to use other pipeline techniques such as snakemake [12] or Nextflow [13]. The source code and installation instructions are publicly available on GitHub under the MIT license, allowing public and private use and modification and distribution of the code (https://github.com/mcollatz/ConsensusPrime, accessed on 22 September 2022). The pipeline is also integrated into a ready-to-use Docker container with all dependencies and necessary programs pre-installed (mcollatz/consensusprime:1.0).

Embedding this pipeline in a Docker container ensures executability on various systems. The ConsensusPrime pipeline is not available as an online service to avoid privacy concerns arising from data transfer to third parties and to ensure usability at all times.

**Author Contributions:** M.C. implemented the pipeline and wrote the manuscript. S.D.B., S.M. and R.E. helped with the conceptual development and definition of the requirements. All authors revised and corrected the various drafts and approved the final version of the manuscript. All authors have read and agreed to the published version of the manuscript.

**Funding:** ADA (13GW0456C: BMBF): Adaptable Decentralized Diagnostics for Veterinary and Human Medicine. The publication of this article was funded by the Open Access Fund of the Leibniz Association.

**Institutional Review Board Statement:** Not applicable.
Informed Consent Statement: Not applicable.

Data Availability Statement: All code and files are available at GitHub (https://github.com/mcollatz/ConsensusPrime).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Watson, J.D.; Crick, F.H. Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. *Nature* 1953, 171, 737–738. [CrossRef] [PubMed]
2. Wang, Y.; Zhao, Y.; Bollas, A.; Wang, Y.; Au, K.F. Nanopore sequencing technology, bioinformatics and applications. *Nat. Biotechnol.* 2021, 39, 1348–1365. [CrossRef] [PubMed]
3. Kleppe, K.; Ohtsuka, E.; Kleppe, R.; Molineux, I.; Khorana, H. Studies on polynucleotides: XCVI. Repair replication of short synthetic DNA’s as catalyzed by DNA polymerases. *J. Mol. Biol.* 1971, 56, 341–361. [CrossRef] [PubMed]
4. Spencer, M. The stereochemistry of deoxyribonucleic acid. II. Hydrogen-bonded pairs of bases. *Acta Crystallogr.* 1959, 12, 66–71. [CrossRef]
5. Mullis, K.B. The unusual origin of the polymerase chain reaction. *Sci. Am.* 1990, 262, 56–65. [CrossRef] [PubMed]
6. Rose, T.M.; Henikoff, J.G.; Henikoff, S. CODEHOP (COnsensus-DEgenerate hybrid oligonucleotide primer) PCR primer design. *Nucleic Acids Res.* 2003, 31, 3763–3766. [CrossRef] [PubMed]
7. Yoon, H. and Leitner, T. PrimerDesign-M: A multiple-alignment based multiple-primer design tool for walking across variable genomes. *Bioinformatics* 2015, 31, 1472–1474. [CrossRef] [PubMed]
8. Katoh, K.; Standley, D.M. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* 2013, 30, 772–780. [CrossRef] [PubMed]
9. Untergasser, A.; Cutcutache, I.; Koressaar, T.; Ye, J.; Faircloth, B.C.; Remm, M.; Rozen, S.G. Primer3—New capabilities and interfaces. *Nucleic Acids Res.* 2012, 40, e115. [CrossRef] [PubMed]
10. Larkin, M.A.; Blackshields, G.; Brown, N.P.; Chenna, R.; McGettigan, P.A.; McWilliam, H.; Valentin, F.; Wallace, I.M.; Wilm, A.; Lopez, R.; et al. Clustal W and Clustal X version 2.0. *Bioinformatics* 2007, 23, 2947–2948. [CrossRef]
11. Monecke, S.; Coombs, G.; Shore, A.C.; Coleman, D.C.; Akpaka, P.; Borg, M.; Chow, H.; Ip, M.; Jatzwauk, L.; Jonas, D.; et al. A field guide to pandemic, epidemic and sporadic clones of methicillin-resistant Staphylococcus aureus. *PLoS ONE* 2011, 6, e17936. [CrossRef] [PubMed]
12. Mölder, F.; Jablonski, K.P.; Letcher, B.; Hall, M.B.; Tomkins-Tinch, C.H.; Sochat, V.; Forster, J.; Lee, S.; Twardziok, S.O.; Kanitz, A. Sustainable data analysis with Snakemake. *F1000Research* 2021, 10, 33. [CrossRef] [PubMed]
13. Di Tommaso, P.; Chatzou, M.; Floden, E.W.; Barja, P.P.; Palumbo, E.; Notredame, C. Nextflow enables reproducible computational workflows. *Nat. Biotechnol.* 2017, 35, 316–319. [CrossRef] [PubMed]