RFLP-Inator: Interactive Web Platform for In Silico Simulation and Complementary Tools of the PCR-RFLP Technique

Kiefer Andre Bedoya Benites and Wilser Andrés García-Quispes

***Abstract*—Polymerase chain reaction - Restriction Fragment Length Polymorphism (PCR-RFLP) is an established molecular biology technique leveraging DNA sequence variability for organism identification, genetic disease detection, biodiversity analysis, etc. Traditional PCR-RFLP requires wet-laboratory procedures that can result in technical errors, procedural challenges, and financial costs. With the aim of providing an accessible and efficient PCR-RFLP technique complement, we introduce RFLP-inator. This is a comprehensive web-based platform developed in R using the package Shiny, which simulates the PCR-RFLP technique, integrates analysis capabilities, and offers complementary tools for both pre- and post-evaluation of in vitro results. We developed the RFLP-inator's algorithm independently and our platform offers seven dynamic tools: RFLP simulator, Pattern identifier, Enzyme selector, RFLP analyzer, Multiplex PCR, Restriction map maker, and Gel plotter. Moreover, the software includes a restriction pattern database of more than 250,000 sequences of the bacterial 16S rRNA gene. We successfully validated the core tools against published research findings. This new platform is open access and user-friendly, offering a valuable resource for researchers, educators, and students specializing in molecular genetics. RFLP-inator not only streamlines RFLP technique application but also supports pedagogical efforts in genetics, illustrating its utility and reliability.**

***Index Terms*—Computer simulation, open educational resources, polymerase chain reaction (PCR), restriction fragment length polymorphism (RFLP), software tools, web design.**

I. Introduction

FINGERPRINTING techniques, such as restriction fragment length polymorphism (RFLP), leverage the natural variability in DNA to distinguish among different sequences. Traditionally, it has long served as a cornerstone technique in molecular biology due to its simplicity and versatility, having a broad range of applications including the detection of genetic disease carriers [1], the characterization of genetic polymorphism [2], [3], the genetic mapping [4], the identification of organisms [5], [6], the analysis of diversity [7], [8], among others.

Received 2 April 2024; revised 26 September 2024; accepted 5 October 2024. Date of publication 8 October 2024; date of current version 10 December 2024. *(Corresponding author: Wilser Andrés García-Quispes.)*

The authors are with the Department of Cellular Biology and Genetics, Faculty of Biological Sciences, National University of San Marcos, Lima 15081, Peru (e-mail: kiefer.bedoya@unmsm.edu.pe; wgarciaq@unmsm.edu.pe).

The software is available for free at https://kodebio.shinyapps.io/RFLP-inator/.

Digital Object Identifier 10.1109/TCBB.2024.3476453

The method primarily consists of three fundamental steps. Initially, the target DNA sequence is amplified by PCR. Subsequently, the PCR products undergo digestion by a specific class of enzymes known as restriction endonucleases (or restrictases); these enzymes, which are predominantly found in bacteria [9], recognize and bind to specific short DNA sequences known as restriction sites. Finally, the resulting DNA fragments are finally visualized by gel electrophoresis. Due to inherent polymorphisms, DNA bands on the gel display a unique distribution, termed as restriction pattern [10]. These distinctive patterns facilitate the differentiation and identification of organisms or genetic variants.

Given the intricate logistics and costs associated with experimental wet-lab methods, an efficient computational alternative for simulating and complementing these procedures has become imperative. This strategy not only streamlines research workflow but also offers pedagogical advantages, especially in settings where physical resources are scarce. The recent COVID-19 pandemic underscored the critical role of informatic resources, revealing the vulnerabilities of relying solely on traditional laboratory techniques. With the widespread closure of research facilities due to quarantine measures, there emerged an accelerated demand for alternative platforms that ensure the continuity of scientific research and educational training [11].

In the field of molecular genetics, several platforms address various aspects of the RFLP technique analysis and simulation. GERM [12] and FragMatch [13] allow the identification of ectomycorrhizal fungi and taxa within mixed samples by matching RFLP and DNA fragment data against established databases, facilitating diversity analysis. Other tools are primarily focused on in silico simulation; for example, PCR-RFLP in silico [14] is a platform specifically developed for the design of PCR-RFLP experiments, while NEBcutter [15] allows the generation of detailed restriction enzyme maps. Furthermore, there exist available resources that do not rely on a graphical user interface, such as RFLPtools [16], an R package which provides a computational suite for RFLP data analysis. Collectively, these platforms offer wide support for genetic analysis, from experimental design through to data interpretation.

In this article, we introduce RFLP-inator, a platform that integrates multiple aspects of the RFLP technique into a cohesive and comprehensive tool. Unlike its counterparts, RFLP-inator not only facilitates detailed RFLP technique analysis but also

TABLE I
COMPARISON OF NEBCUTTER, RFLP-TOOLS, IN SILICO PCR-RFLP AND RFLP-INATOR TOOLS

Features	NEBcutter v3.0	RFLP-tools	*In silico* PCR-RFLP	RFLP-inator
Supported environments	Web browser	R Console	Web browser	Web browser
Programming language	C	R	PHP	R
User interface	Yes	No	Yes	Yes
PCR simulation	No	No	Yes	Yes
Restriction enzyme digestion simulation	Yes	No	Yes	Yes
Methods for handling restriction site overlapping	No	N/A	No	Yes
Provide restriction fragment sequences	No	N/A	No	Yes
Compare restriction pattern similarities	No	Yes	No	Yes
Analyze *in vitro* RFLP data	No	Yes	No	Yes
Generate restriction maps	Yes	No	No	Yes
Identify enzymes which generate unique restriction patterns across sequences	No	No	No	Yes

extends its utility to the simulation of multiplex scenarios. It encompasses additional complementary tools that address the broader needs surrounding the RFLP technique. This integration creates a seamless transition fromsimulation to analysis, providing a comprehensive suite that enhances the efficiency of genetic examination. Table I summarizes the capabilities and limitations of some RFLP-related tools in comparison to RFLP-inator. Moreover, RFLP-inator incorporates a proprietary algorithm for amplification, digestion, and restriction pattern comparison. This algorithm enables the simulation of specific scenarios such as non-specific primer hybridization, overlap within restriction enzyme recognition sequences, and the comparison of restriction patterns with a different number of bands that may be indistinguishable in vitro. Designed to be both user-friendly and freely accessible, RFLP-inator ensures that researchers or students can easily engage with its functionalities, while also contributing to the democratization of access to genetic analysis tools. These capabilities make RFLP-inator a unique tool, offering great versatility in the analysis and simulation of RFLP data.

II. SOFTWARE DEVELOPMENT

A. Development Framework and Data Sources

The RFLP-inator web tool was developed using the R programming language [17] and a web interface development package called Shiny [18]. This two-tiered development approach involved coding for PCR-RFLP technique simulation, alongside a user-friendly graphical interface. The tool makes available a selection of 344 different restriction site patterns corresponding to 1501 type II restriction enzymes sourced from the REBASE database [19]. The data was downloaded from http://rebase.neb.com/rebase/rebase.f7.html and filtered to exclude enzymes with nonspecific cleavage sites or recognition sequences shorter than three nucleotides. Following the application of these filters, 1501 restriction enzymes remained, including type IIP enzymes such as *Eco*RI and *Bgl*I, type IIS enzymes like *Fok*I and *Bin*I, and type IIG enzymes such as *Acu*I and *Bcg*I. In addition, as a complement to the OTU identifier tool, Bacterial identifier tool required bacterial 16S rRNA gene sequences, which were sourced from the Ribosomal Database Project [20]. Sequences underwent filtering, amplification, and digestion as detailed below.

RFLP-inator is currently hosted on Shinyapps.io (www.shinyapps.io), a cloud-based platform that facilitates web applications built with R, ensuring broad accessibility and ease of use across a wide range of devices without requiring local installations. In terms of performance, our software is subject to the resource constraints of the server.

B. In Silico PCR-RFLP Simulation

1) PCR Amplification Simulation: The initial phase in performing PCR-RFLP in vitro entails the generation of multiple copies of the target sequence (DNA amplification). However, for process simulation, only the amplicon sequence is needed, as the restriction pattern solely represents amplified DNA. The software identifies hybridization sites for the forward and reverse primers, considering only exact matches between the primer sequences and the target DNA in order to ensure specificity. Once recognized, the sequence between the first hybridized nucleotide with the forward primer and the last hybridized nucleotide with the reverse primer's complementary sequence is delimited and saved.

In the scenario of non-specific primer hybridization (hybridization across multiple sites), the amplification simulation will consider each of the potential fragments generated. Since PCR works with double-stranded DNA, primers can hybridize to either strand due to sequence complementarity; however, the simulation of amplification focuses solely on the forward strand. The software algorithm, therefore, searches for both the primer sequences and their complements in the DNA to ensure accurate simulation. As shown in Fig. 1, each amplicon starts from the binding site of a hybridized primer on the complementary strand (3'→5' orientation) and extends to the binding site of each hybridized primer on the forward strand (5'→3' orientation).

2) Enzymatic Digestion Simulation: Following PCR amplification, the resulting amplicon undergoes digestion by restriction enzymes. To accurately predict the products of DNA sequence digestion, we developed a novel algorithm. Considering potential sticky ends (single-stranded DNA extremes or ssDNA) in the restriction product, the initial step transforms DNA sequences into a coded format of numbers and letters, where numbers

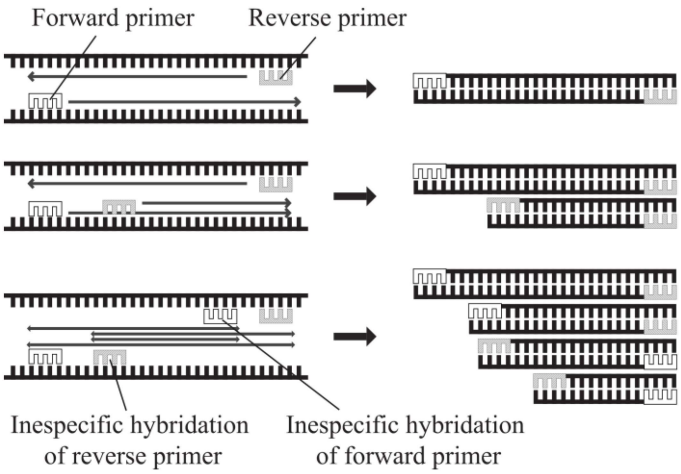

Fig. 1. Amplification products by in silico PCR in different instances. Comparison of PCR amplification with specific hybridization of primers versus amplification with nonspecific hybridization of primers. The figure illustrates the potential for generating multiple amplicons due to nonspecific interactions.

represent double-stranded DNA (dsDNA), uppercase letters indicate overhang in the forward strand and lowercase letters represent overhang in the complementary strand. This encoding system allows for the efficient discernment of the nature of DNA products post-digestion. Restriction endonucleases undergo the same encoding process.

Subsequent steps involve identifying the hybridization sites of restriction enzymes on the target DNA, requiring exact matches between sequences. Once determined, a caret symbol (^), an asterisk ($*$) or a vertical bar (|) is incorporated in sequences: the caret symbol represents a cut in the forward strand; the asterisk, in the complementary strand; and vertical bar, in both strands simultaneously (blunt end). Finally, the software locates all symbols in the sequence and sets the limits as appropriate. Given that some enzymes recognize degenerate DNA sequences, all variable restriction site patterns are considered in the matching process. Furthermore, the algorithm takes into account the complement of each restriction site sequence, as not all enzymes recognize palindromic sequences; their restriction sites can also be detected on the complementary strand of the target DNA, thereby ensuring comprehensive detection of all standard enzyme hybridizations.

The described process represents the most straightforward scenario; however, there are instances where a restriction site pattern overlaps with itself within the same DNA, as sketched in Fig. 2. In such situations, the program operates as described: initially, all non-overlapping restriction sites are cleaved; then the overlapping sites are identified, and the sequence is multiplied by the number of unique enzyme binding possibilities; subsequently, each potential binding site is independently cleaved; finally, the products follow the same procedure, recursively. This methodology is applied because once the endonuclease cleaves one of the overlapping restriction sites, the remaining site will not be identified in the product. Therefore, the designed algorithm enables simulated digestion under any potential circumstance, including degenerate, non-palindromic, or multiple overlapping restriction site patterns within the DNA sequence.

3) Electrophoresis Simulation: Electrophoresis was simulated using base R commands for plotting (low-level graphic), considering four parameters: (1) molecular weights of the bands, (2) the ladder (molecular weight marker), (3) the ladder's positions on the gel, and (4) the lane names.

C. Pattern Identification

The comparing of different restriction patterns allows for the identification of identical patterns, assuming they were generated by the same restriction enzyme. This is the process followed by Pattern identifier, which determines the potential organism that generated the given restriction pattern based on the comparison of band weights. To perform this, the software requires: (1) a restriction patterns database from a group of organisms, which can be produced by RFLP-inator by providing a set of gene sequences (in FASTA format); (2) the molecular weights of the bands generated after in vitro digestion and electrophoresis; (3) the enzyme that was used; and (4) a threshold representing the acceptable error as a measure of uncertainty. The restriction pattern database must be submitted in CSV format, comprising all restriction patterns resulting from the digestion of each of the uploaded sequences by all restriction enzymes featured in the tool. This file can be generated directly by RFLP-inator.

The unknown input restriction pattern is compared against all patterns in the database generated by the same endonuclease. For this purpose, a similarity score is calculated as follows:

If patterns have a same number of bands ($n_a = n_b = n$):

$$s = \left(\frac{1}{n}\right) \sum_{i=1}^{n} |\log bp_{a_i} - \log bp_{b_i}| \tag{1}$$

where:

- bp_a: molecular weights (in base pairs) of the "a" restriction pattern
- bp_b: molecular weights (in base pairs) of the "b" restriction pattern

If patterns have a different number of bands ($n_a > n_b$):

$$s = \left(\frac{1}{n_a}\right) \sum_{i=1}^{n_a} |\log bp_{a_i} - \log bp_{b'_i}| \tag{2}$$

where:

- bp_a: molecular weights (in base pairs) of the "a" restriction pattern
- $bp_{b'}$: molecular weights (in base pairs) constructed from "b" by duplicating selected bands

According to (1), a lower score indicates higher similarity between the restriction patterns. In instances where the restriction patterns feature an unequal number of bands, the similarity score is calculated in a similar way but standardizing this number (2). Both patterns are sorted in descending order according to their molecular weight, then missing bands, in the pattern with fewer bands, are supplemented by duplicating those bands from that same pattern that exhibit the smallest logarithmic size difference relative to any bands in the more extensive pattern. This process is iterated until both patterns have an equal number of bands.

Pattern identification can be performed with any database input by the user. Furthermore, RFLP-inator includes a preloaded

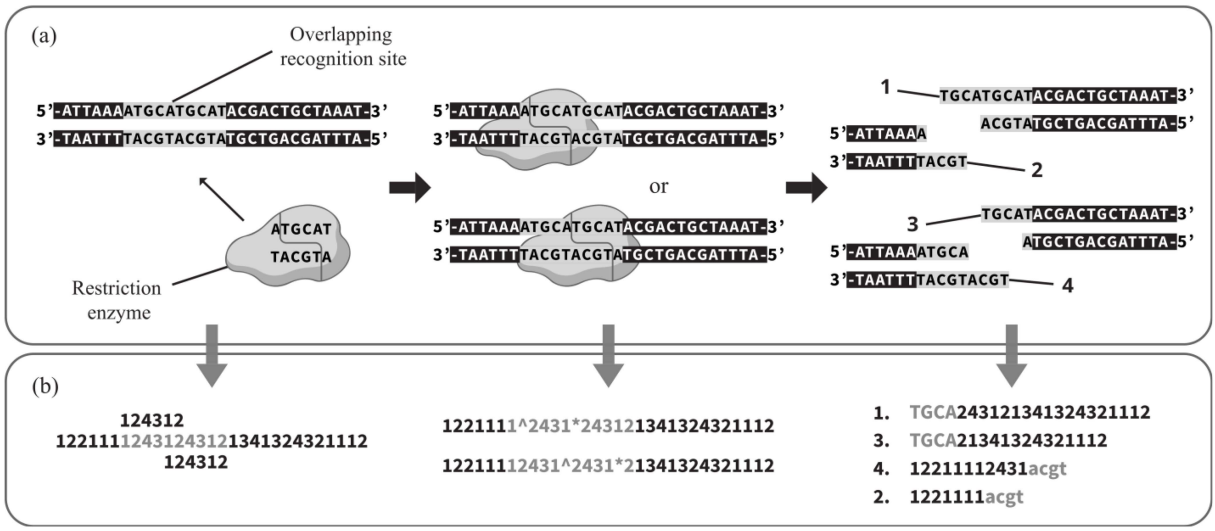

Fig. 2. Simulation of Enzymatic Digestion. (a) Molecular representation of enzyme cleavage at overlapping restriction sites in the DNA. Binding at one site renders the other site inaccessible. (b) Algorithmic encoding of the digestion products.

database of restriction patterns generated by the amplification and digestion of a curated set of 16S rRNA gene sequences sourced from the Ribosomal Database Project database. The sequences were in silico amplified using two pairs of universal 16S rRNA primers: F27 (5'- AGAGTTTGATCMTGGCTCAG-3') and R1492 (5'- TACGGYTACCTTGTTACGACTT -3'), as described by Heuer et al. [21], and another set, F27 and R926 (5'- CCGYCAATTYMTTTRAGTTT-3'), as detailed by Quince et al. [22].

D. Selection of Enzymes That Generate Unique Pattern

By employing the similarity score and a defined threshold, the tool allows for identification of enzymes that generate different patterns for a given sequence. The process requires a set of DNA sequences in FASTA format, which are digested in silico with all restriction enzymes in the database. Only enzymes that generate a distinct pattern (by threshold) between the selected sequence and the remaining ones will be included in the results. Additionally, it is possible to identify which enzymes generate distinct patterns across all sequences. This is achieved by comparing, per enzyme, each restriction pattern against all others. When any comparison results in a similarity score below the predefined threshold, that enzyme is discarded.

E. In Vitro RFLP Results Analysis

The analysis of in vitro PCR-RFLP results in a molecular fingerprinting study entails creating a distance matrix, utilized for generating graphical representations such as: (1) dendrograms, via various hierarchical clustering methods; (2) heatmaps, for visualizing sample similarity; and (3) multidimensional scaling, to spatially represent sample relationships. Each of these plots is built using the distance score provided by the RFLPtools package [16], a specialized R package designed for the analysis of RFLP data. Additionally, the MKmisc and lattice packages were employed to further support the visualization and interpretation of the results [23], [24].

F. User Interface

The user interface was developed using the Shiny package [18] with additional support from the bs4Dash package [25], enabling the execution of developed functions in a graphical environment. RFLP-inator features seven tools to meet user requirements: (1) RFLP simulator, (2) Pattern identifier (which includes OTU identifier and Bacterial identifier), (3) Enzyme selector, (4) RFLP analyzer, (5) Multiplex PCR, (6) Restriction map maker, and (7) Gel plotter. The interface is equipped with several usability features, including a dark mode option, which helps reduce eye strain during extended sessions. In addition, tooltips are integrated throughout the interface, providing immediate, context-sensitive guidance and explanations for key features and inputs.

III. RFLP-Inator's Tools

The results of the work include each of the tools offered by the RFLP-inator web program (https://kodebio.shinyapps.io/RFLP-inator/). Fig. 3 shows a picture of the tool's interface.

A. RFLP Simulator

RFLP simulator is a core component of the toolset, designed to perform in silico simulations of PCR-RFLP. This module generates a simulated agarose gel image representing the DNA fragments produced post-digestion for a given target sequence. Users can digest one sequence with multiple enzymes or multiple sequences with a single enzyme. The gel image is accompanied by a data table detailing the lengths, in base pairs, of the resultant fragments for each electrophoretic lane. Additionally, the sequence of each fragment product can be visualized.

RFLP Simulator requires four categories of input data:

1) *Sequence:* The DNA sequence targeted for PCR-RFLP simulation must be provided by the user. This can be achieved by either inputting NCBI accession codes,

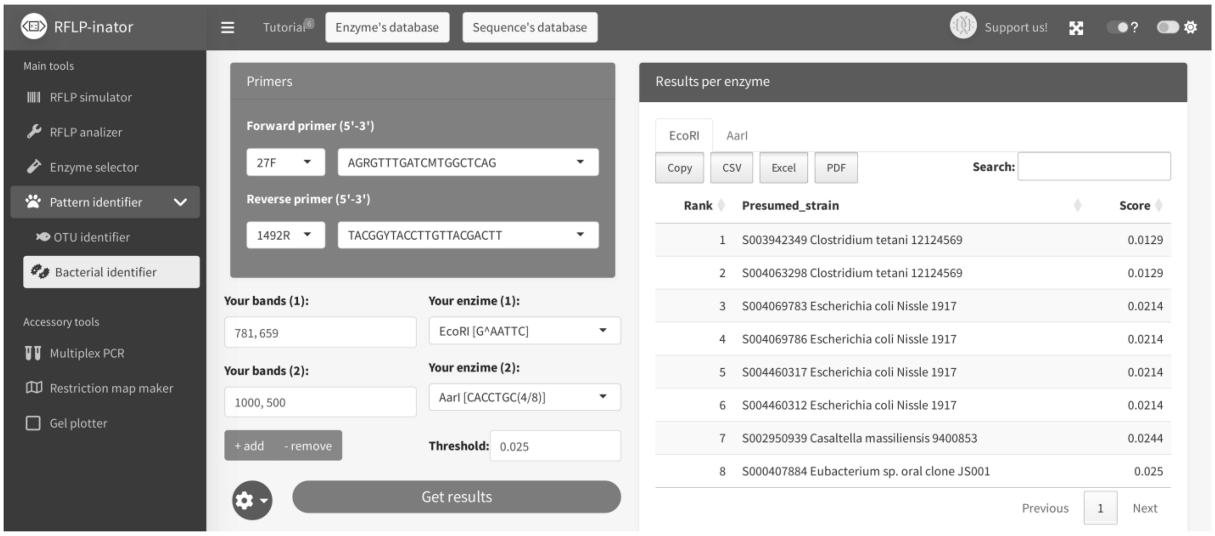

Fig. 3. RFLP-inator interface. The image displays Bacterial identifier interface as an example.

pasting the sequence(s) into a dedicated text input field, or uploading a FASTA-formatted file.

2) *Primers:* The user specifies the forward and reverse primers (5'→3') for the target sequence amplification. Amplification is optional in case the user inputs amplified sequences.
3) *Enzyme Options:* In this section, users can select one or more restriction enzymes from the tool's built-in database to perform simulated DNA digestion. It is also possible to input a customized restriction site following the REBASE database format.
4) *Ladder Options:* The tool includes various molecular ladders for users to select from. Additionally, it is possible to modify the weights of the molecular marker.

B. Pattern Identifier

As a result of the in vitro application of the PCR-RFLP technique, the molecular weights of the fragments obtained from the digestion of a DNA with any enzyme can be determined. The Pattern identifier tool uses these results to predict which possible sequence, from those present in a database, could have generated the entered restriction pattern. Users are required to input (1) the length, in base pairs, of each experimentally obtained fragment, (2) the restriction enzyme used in the digestion process, and (3) the forward and reverse primers employed for DNA amplification. A threshold value is also set to filter the results, ensuring that only sequences with a similarity score below the threshold will be included.

The tool generates result tables that list potential sequences identified independently for each restriction enzyme. An additional feature allows for a comparative analysis across the tables and includes an electrophoretic gel schematic to provide visual context. This gel comprises the lanes for the ladder, the input fragments, and a simulated digestion of each sequence with a similarity score below the set threshold. The sequence with the lowest similarity score is selected for each enzyme, and a subsequent digestion simulation is conducted to produce another electrophoretic gel diagram. In this instance, the lanes correspond to the fragments produced by the specific restriction, providing further analytical depth and enhancing the reliability of the PCR-RFLP analysis by cross-referencing experimental data with computational predictions.

Pattern identifier offers two modules:

1) *Bacterial Identifier:* This tool utilizes a preloaded 16S rRNA gene restriction pattern database constructed from all the sequences present in the Ribosomal Database Project and REBASE's enzymes with the objective of recognizing possible bacterial strains that have generated the 16S rRNA restriction pattern. More than 250000 strains were in silico amplified and digested. Users can include unculturable bacteria in their search.
2) *OTU Identifier:* This approach requires users to upload their own sequence database in FASTA format. From this input, a specialized sub-database of restriction patterns is constructed, which can be downloaded and subsequently utilized for identification.

In vitro electrophoresis can present certain limitations, such as the superposition of two bands with nearly identical molecular weights, rendering them indistinguishable. Should this occur, and the user inputs a single fragment into Pattern identifier when there are actually two, the results could be systematically biased. To avoid this, an option for comparing restriction patterns with different number of bands can be activated. Upon activation, the algorithm for similarity scoring is adjusted, as detailed in Section 2.3, thereby overcoming this limitation. Another challenge arises from the potential loss of small fragments in conventional electrophoresis. To address this, users have the option to set a minimum band size for consideration in similarity scoring, ensuring only bands larger than the specified size are included.

C. Enzyme Selector

The Enzyme selector tool is designed to identify restriction enzymes that yield unique restriction patterns among a set of sequences. Using a comparative algorithm, the tool analyzes

the restriction patterns of each sequence with a specific enzyme and verifies whether the resulting patterns are mutually distinct. Should the patterns be unique, the enzyme is shortlisted in the output. This process is systematically executed for every enzyme in the tool's database. Similarly, Enzyme selector also allows for the differentiation of a single sequence within a set. In such instances, only enzymes that generate a unique restriction pattern for the selected sequence will be included in the output, regardless of whether the remaining patterns are identical to one another.

This tool requires four types of input data:

1) *Sequences:* Users must upload a set of at least two sequences. Subsequently, they must select from two analysis choices: "Different patterns among all sequences" or "Different pattern for one sequence".
2) *Primers:* User specifies the forward and reverse primers. If any of the primers do not align with the entered sequence, the tool will display a warning indicating the failure of amplification. An option to disable the simulation of amplification is also available.
3) *Threshold:* The threshold parameter establishes a minimum score that determines whether restriction patterns are sufficiently distinct from one another. If the similarity score between two patterns falls below the set threshold, they are considered identical, thereby excluding the corresponding enzyme from the output.
4) *Enzyme List:* Enzymes that meet the aforementioned criteria are displayed in a selection output. This allows users to visualize comparative restriction patterns in simulated electrophoretic gels as part of the output results.

D. RFLP Analyzer

The RFLP Analyzer is tailored to facilitate the analysis of data from in vitro PCR-RFLP results in molecular fingerprinting studies. To initiate the analysis, users must upload a table in CSV format, listing the molecular weights of each band from an electrophoretic gel obtained in vitro, as stipulated by the RFLPtools package. Once inputted, it grants the ability to generate and show a distance matrix for sample comparisons. RFLP Analyzer accommodates the generation of dendrograms utilizing various hierarchical clustering methodologies, such as Euclidean, Maximum, Manhattan, Canberra, Minkowski, and Pearson correlation distance.

Users can adjust the Level of Detail (LOD) parameter, which sets the threshold for band detection sensitivity. The p-value in the Minkowski metric can also be modified to adjust its power. Beyond dendrograms, the RFLP Analyzer tool provides additional visualizations such as (Dis-)Similarity Plots for detailed similarity assessments and Multidimensional Scaling for spatial representation of genetic distances. These features collectively serve the demands of RFLP data analysis.

E. Multiplex PCR

This tool was designed to simulate the multiplex polymerase chain reaction (multiplex PCR) process, wherein multiple primers and target sequences can be inputted according to the user's research needs. The simulation aims to predict the amplicon size of a multiplex PCR and displays the results in a table listing the molecular weights. Multiplex PCR also generates a virtual electrophoretic gel, where each input sequence is represented in a distinct lane, allowing for clear visualization of the resultant bands.

F. Restriction Map Maker

Restriction map maker generates restriction maps based on an input DNA sequence using all the restriction enzymes included in the tool's database. A restriction map is a graphical layout of known restriction sites within the DNA sequence [26]. The generated graph serves as a visual guide to identify enzymes capable of cleaving the sequence at single or multiple locations.

The input requirements consist solely of the DNA sequence and, optionally, associated primers. After the graph is generated, users have the option to modify specific parameters to meet their experimental needs. These adjustable settings include the maximum number of cleavage sites, sequence length in base pairs, the orientation of enzyme display on the graph, enzyme clustering on the map, and a selection input to specify which enzymes will be displayed.

G. Gel Plotter

The Gel Plotter feature is a component of RFLP-inator designed to offer users the capability to customize electrophoretic gels according to their specific requirements. The function enables the adjustment of various parameters, including band positioning, the number of lanes, lane labels, weights of the ladder's bands, etc.

IV. Major Components Validation

The functionality of RFLP-inator's primary tools was validated by comparing in vitro research outcomes with those generated by the web-based tool. In 2016, Mandakovic et al. performed a selection of restriction enzymes for the specific detection of *Piscirickettsia salmonis* using 16S rRNA PCR-RFLP (referred to in the paper as 16S rDNA PCR-RFLP) [27]. They identified enzymes such as *Bsa*AI, *Oli*I, and *Pma*CI (type II) as effective for the pathogen's detection. Using the Bacterial identifier module, the molecular weights of the bands produced by each enzyme, as reported in the paper, were inputted to determine which strains could have produced the observed patterns. The results were as expected, as shown in Fig. 4(a), RFLP-inator identifies strains of *Piscirickettsia salmonis* as potential candidates.

Further validation of RFLP-inator was conducted using data from the same study. A set of sequences (KU204881-KU204892) obtained from the NCBI database, previously analyzed by Mandakovic et al. using PCR-RFLP with the restriction enzyme *Pma*CI, was processed through RFLP simulator. The simulation results were consistent with their in vitro findings (Fig. 4(b) and middle panel of Fig. 2 in Mandakovic et al. article [27]).

Additionally, Figueras, Collado, and Garro (2008) presented a 16S rRNA PCR-RFLP method (referred to in the paper as 16S rDNA-RFLP) for discriminating *Arcobacter* species [28].

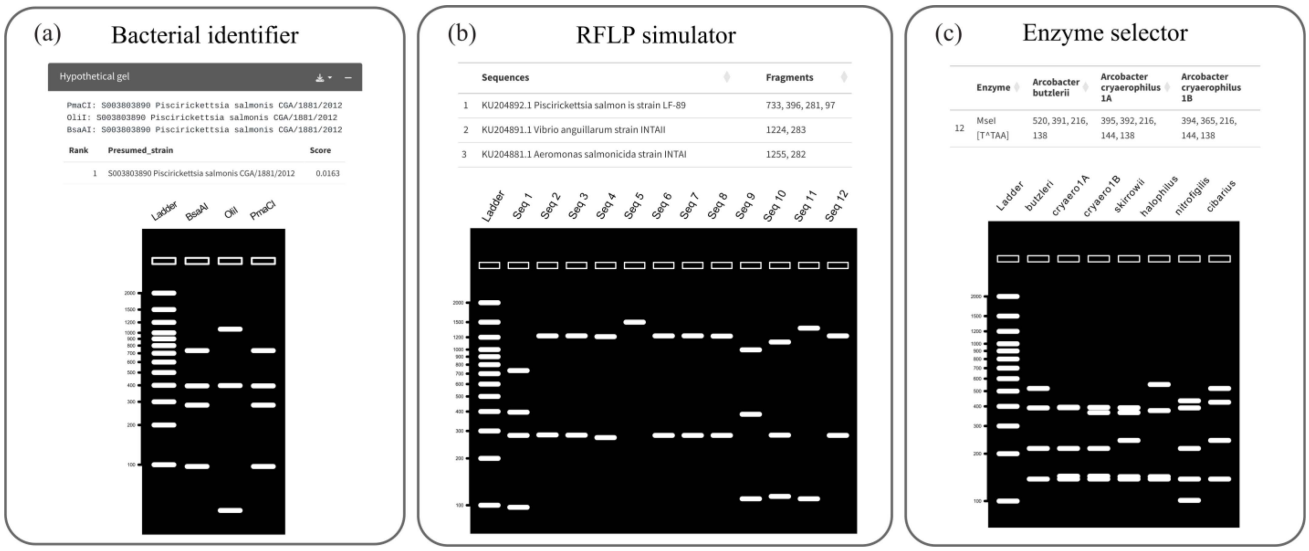

Fig. 4. Validation of RFLP-inator tools using in vitro data. (a) Using the PCR-RFLP results from Mandakovic et al. [27] for *Piscirickettsia salmonis*, Bacterial identifier accurately recognized the restriction pattern. (b) Based on the same study, RFLP simulator successfully replicated in vitro restriction patterns for various 16S rRNA gene sequences obtained from NCBI. (c) Enzyme selector identified *Mse*I as an optimal enzyme for differentiating *Arcobacter* species, as proposed by Figueras et al. [28]. Note: Tables within the figure are partially displayed.

The enzyme *Mse*I (type II) was highlighted as the only one among those studied that yielded distinct restriction patterns among 16S rRNA gene sequences of the species *A. butzleri*, *A. cryaerophilus* 1A, *A. cryaerophilus* 1B, *A. skirrowii, A. halophilus, A. nitrofigilis,* and *A. cibarius* (GenBank accession codes: L14626, L14624, AY314755, L14625, AF513455, L14627, and AJ607391, respectively). The Enzyme Selector module was employed to corroborate the findings reported in the paper. For this validation, we set a threshold of 0.01, and the sequences, in FASTA format, were uploaded. We excluded fragments with lengths shorter than 100 bp, in accordance with the original paper. The program selected *Mse*I as one of the enzymes generating unique patterns among the sequences, as illustrated in Fig. 4(c).

V. Limitations

While RFLP-inator provides a comprehensive platform for in silico PCR-RFLP simulations, certain limitations must be acknowledged. One key limitation is the software's inability to consider methylation sites within DNA sequences. The restriction enzymes in the tool will cleave at restriction sites irrespective of methylation status, which can lead to inaccurate results in the analysis of restriction patterns. In fact, methylation of the target DNA can inhibit enzyme activity by preventing the enzymes from recognizing and cleaving the restriction site [29]; moreover, some enzymes, such as BisI, cleave specifically methylated DNA [30]. Furthermore, the simulation of amplification requires an exact match between the primer and the DNA sequence to proceed. Nonetheless, in vitro, mismatches still result in hybridization. To address this, RFLP-inator includes the option to disable the amplification step when users provide pre-amplified sequences (amplicons). Another limitation is that the software assumes linear DNA in its simulations, which may create challenges when analyzing more complex structures such as circular DNA or highly repetitive regions, as these can influence both enzyme cleaving efficiency and electrophoresis migration patterns. Moreover, as with any simulation tool, RFLP-inator operates under idealized conditions and does not account for stochastic events commonly encountered in laboratory experiments, such as incomplete digestion, variability in enzyme efficiency, or differences in band migration and staining intensity during electrophoresis, all of which may affect the outcomes. Future updates to RFLP-inator will focus on addressing these limitations, potentially enhancing the accuracy and applicability of its simulations.

VI. Conclusion

RFLP-inator offers a comprehensive platform that integrates simulation, analysis, and complementary tools for Restriction Fragment Length Polymorphism (RFLP) technique. The tool's proprietary algorithm for amplification, digestion, and restriction pattern comparison is able to simulate complex scenarios, such as non-specific primer hybridization or overlapping restriction enzyme recognition sequences, distinguishing it from other tools in the field. Its user-friendly interface and free accessibility ensure that RFLP-inator is a viable option for use by researchers, professors, and biotechnology students. Each of the designed tools functions according to the objectives proposed for their development. Moreover, the successful validation of RFLP-inator's functionalities through comparison with in vitro research outcomes highlights its reliability and accuracy in simulating and analyzing RFLP data. The R programming language made it possible to translate the proposed workflow into lines of code; likewise, the Shiny package facilitated the development of a web interface that allows the tool to have a user-friendly interface. On balance, RFLP-inator embodies a complete, comprehensive, accessible, and reliable solution for diverse needs related to the RFLP technique and more.

ACKNOWLEDGMENT

The authors express sincere gratitude to Jorge Ramirez and his team at the Genomics and Bioinformatics for Biodiversity Laboratory, Department of Biological Sciences, National University of San Marcos, for generously allowing us access to their server. This access was crucial for constructing the database of 16S rRNA bacterial restriction fragments. The authors also thank to Leyda Cabrera for her assistance with the English language editing.

REFERENCES

[1] V. Permatasari, M. I. Saleh, and S. Nita, "Estrogen receptor alpha (ERα) Pvu II 397 T/C related genotypes and alleles are associated with higher susceptibilities of endometriosis," *J. Phys.: Conf. Ser.*, vol. 1485, no. 1, 2020, Art. no. 12011, doi: 10.1088/1742-6596/1485/1/012011.

[2] H.-J. Paek, Z.-Y. Li, B.-H. Quan, and X.-J. Yin, "Application of PCR-RFLP for quick identification of MSTN mutants in MSTN mutant pig breeding," *Animal Biotechnol.*, vol. 34, no. 7, pp. 2231–2239, Dec. 2023, doi: 10.1080/10495398.2022.2083628.

[3] A. R. R. Tolee, E. Olga, and C. Ekaterina, "Identification of CLPG gene polymorphism using PCR-RFLP of Iraq and Belarus population sheep breeds," *Gene Rep.*, vol. 22, 2021, Art. no. 100974, doi: 10.1016/j.genrep.2020.100974.

[4] T. Begna and H. Yesuf, "Genetic mapping in crop plants," *Open J. Plant Sci.*, vol. 6, no. 1, pp. 19–26, 2021, doi: 10.17352/ojps.000028.

[5] T. Yuhara, H. Ohtsuki, and J. Urabe, "A simple method for species identification of the ghost crabs using PCR-RFLP," *Plankton Benthos Res.*, vol. 18, no. 2, pp. 106–109, 2023, doi: 10.3800/pbr.18.106.

[6] S. Zhu et al., "Genetic identification of medicinally used Salacia species by nrDNA ITS sequences and a PCR-RFLP assay for authentication of Salacia-related health foods," *J Ethnopharmacol*, vol. 274, 2021, Art. no. 113909, doi: 10.1016/j.jep.2021.113909.

[7] D. Zhao et al., "Diversity analysis of bacterial community compositions in sediments of urban lakes by terminal restriction fragment length polymorphism (T-RFLP)," *World J. Microbiol. Biotechnol.*, vol. 28, no. 11, pp. 3159–3170, 2012, doi: 10.1007/s11274-012-1126-y.

[8] S. V. Ramanaiah, C. M. Cordas, S. C. Matias, M. V. Reddy, J. H. Leitão, and L. P. Fonseca, "Bioelectricity generation using long-term operated biocathode: RFLP based microbial diversity analysis," *Biotechnol. Rep.s*, vol. 32, 2021, Art. no. e00693, doi: 10.1016/j.btre.2021.e00693.

[9] B. P. Anton and R. J. Roberts, "Beyond Restriction Modification: Epigenomic Roles of DNA Methylation in Prokaryotes," *Annu. Rev. Microbiol.*, vol. 75, no. 75, 2021, pp. 129–149, 2021, doi: 10.1146/annurev-micro-040521-035040.

[10] R. C. Williams, "Restriction fragment length polymorphism (RFLP)," *Amer. J. Biol. Anthropol.*, vol. 32, no. S10, pp. 159–184, Jan. 1989, doi: 10.1002/ajpa.1330320508.

[11] S. Ray and S. Srivastava, "Virtualization of science education: A lesson from the COVID-19 pandemic," vol. 11, pp. 77–80, 2020, doi: 10.1007/s42485-020-00038-7.

[12] I. A. Dickie, P. G. Avis, D. J. McLaughlin, and P. B. Reich, "Good-Enough RFLP Matcher (GERM) program," *Mycorrhiza*, vol. 13, no. 3, pp. 171–172, 2003, doi: 10.1007/s00572-003-0225-x.

[13] T. A. Saari, S. K. Saari, C. D. Campbell, I. J. Alexander, and I. C. Anderson, "FragMatch—A program for the analysis of DNA fragment data," *Mycorrhiza*, vol. 17, no. 2, pp. 133–136, 2007, doi: 10.1007/s00572-006-0102-5.

[14] R. M. San Millán, I. Martínez-Ballesteros, A. Rementeria, J. Garaizar, and J. Bikandi, "Online exercise for the design and simulation of PCR and PCR-RFLP experiments," *BMC Res. Notes*, vol. 6, no. 1, 2013, Art. no. 513, doi: 10.1186/1756-0500-6-513.

[15] T. Vincze, J. Posfai, and R. J. Roberts, "NEBcutter: A program to cleave DNA with restriction enzymes," *Nucleic Acids Res.*, vol. 31, no. 13, pp. 3688–3691, Jul. 2003, doi: 10.1093/nar/gkg526.

[16] F. Flessa, A. Kehl, and M. Kohl, "Analysing diversity and community structures using PCR-RFLP: A new software application," *Mol. Ecol. Resour.*, vol. 13, no. 4, pp. 726–733, Jul. 2013, doi: 10.1111/1755-0998.12094.

[17] W. Chang et al., "shiny: Web Application Framework for R," 2024. [Online]. Available: https://CRAN.R-project.org/package=shiny

[18] RStudio, Inc, "shiny: Easy web applications in R," 2014.

[19] R. J. Roberts, T. Vincze, J. Posfai, and D. Macelis, "REBASE—A database for DNA restriction and modification: Enzymes, genes and genomes," *Nucleic Acids Res.*, vol. 43, no. D1, pp. D298–D299, Jan. 2015, doi: 10.1093/NAR/GKU1046.

[20] J. R. Cole et al., "Ribosomal database project: Data and tools for high throughput rRNA analysis," *Nucleic Acids Res.*, vol. 42, no. D1, pp. D633–D642, Jan. 2014, doi: 10.1093/NAR/GKT1244.

[21] H. Heuer, M. Krsek, P. Baker, K. Smalla, and E. M. H. Wellington, "Analysis of actinomycete communities by specific amplification of genes encoding 16S rRNA and gel-electrophoretic separation in denaturing gradients," *Appl. Environ. Microbiol.*, vol. 63, no. 8, 1997, Art. no. 3233, doi: 10.1128/AEM.63.8.3233-3241.1997.

[22] C. Quince, A. Lanzen, R. J. Davenport, and P. J. Turnbaugh, "Removing Noise From Pyrosequenced Amplicons," *BMC Bioinf.*, vol. 12, no. 1, 2011, Art. no. 38, doi: 10.1186/1471-2105-12-38.

[23] D. Sarkar, *Lattice: Multivariate Data Visualization with R*. New York, NY, USA: Springer, 2008. [Online]. Available: http://lmdvr.r-forge.r-project.org

[24] M. Kohl, "{MKmisc}: Miscellaneous functions from {M}. {K}ohl," 2022. [Online]. Available: https://github.com/stamats/MKmisc

[25] D. Granjon, "bs4Dash: A 'Bootstrap 4' Version of 'shinydashboard,"' 2024. [Online]. Available: https://cran.r-project.org/package=bs4Dash

[26] H. O. Smith and M. L. Birnstiel, "A simple method for DNA restriction site mapping," *Nucleic Acids Res.*, vol. 3, no. 9, pp. 2387–2398, Sep. 1976, doi: 10.1093/nar/3.9.2387.

[27] D. Mandakovic et al., "Genomic-based restriction enzyme selection for specific detection of Piscirickettsia salmonis by 16S rDNA PCR-RFLP," *Front Microbiol.*, vol. 7, pp. 1–12, 2016, doi: 10.3389/fmicb.2016.00643.

[28] M. J. Figueras, L. Soler, M. R. Chacón, J. Guarro, and A. J. Martínez-Murcia, "Extended method for discrimination of Aeromonas spp. by 16S rDNA RFLP analysis," *Int. J. Systematic Evol. Microbiol.*, vol. 50, no. 6, pp. 2069–2073, Nov. 2000, doi: 10.1099/00207713-50-6-2069.

[29] P. W. Laird, "Principles and challenges of genome-wide DNA methylation analysis," *Nature Rev. Genet.*, vol. 11, pp. 191–203, Feb. 2010, doi: 10.1038/nrg2732.

[30] S. Xu, P. Klein, S. Kh. Degtyarev, and R. J. Roberts, "Expression and purification of the modification-dependent restriction enzyme BisI and its homologous enzymes," *Sci. Rep.*, vol. 6, 2016, doi: 10.1038/srep28579.

Kiefer Andre Bedoya Benites received the bachelor's degree in genetics and biotechnology from the Universidad Nacional Mayor de San Marcos (Lima, Peru), in 2023. From 2019 to 2022, he was a member of the research group "Genetics of Metabolic Diseases" (GENMETAB) where he developed his skills in molecular techniques and programming. Since 2023, he has been serving as a bioinformatics technician in the Laboratory of Genomics and Bioinformatics for Biodiversity, undertaking data mining and genomic studies across various species. His research primarily focuses on data analysis, software and pipeline development, and bioinformatics.

Wilser Andrés García-Quispes received the bachelor of science degree in biology with a specialization in Cell Biology and Genetics from Universidad Nacional Mayor de San Marcos, Lima, Peru, in 2005, the postgraduate degree with Universitat Autònoma de Barcelona, Barcelona, Spain, and the PhD degree in advanced genetics in 2012. His research interests include genotoxicity, genetic association studies (SNPs-disease), and human and cancer cytogenetics. He currently serves as an assistant professor with the Faculty of Biological Sciences, Universidad Nacional Mayor de San Marcos and is a member of the research group "Genetics of Metabolic Diseases" (GENMETAB). His teaching activities encompass genetics and bioinformatics at both the undergraduate and graduate levels, emphasizing the use of the R programming language for genetic association studies and machine learning.