ThermoPlex: An Automated Design Tool for Target-specific Multiplex PCR Primers based on DNA Thermodynamics

Altair Agmata[1,2*], Kevin Labrador[2], Joseph Dominic Palermo[2,3], and Maria Josefa Pante[2]

[1]Department of Chemical Engineering, University of the Philippines, Diliman, Quezon City, 1101, Philippines
[2]The Marine Science Institute, University of the Philippines, Diliman, Quezon City, 1101, Philippines
[3]Institute of Environmental Science and Meteorology, University of the Philippines, Diliman, Quezon City, 1101, Philippines

December 29, 2024

Abstract

Multiplex PCR-based assays are indispensable platforms for rapid and cost-effective DNA-based multi-target detection. The success of such an assay highly depends on the accurate design of oligonucleotide primers, arguably its most vital component. In this study, the ThermoPlex design tool is introduced, offering an automated design pipeline for target-specific multiplex PCR primers motivated by DNA thermodynamics. From a sequence alignment of all relevant target and non-target sequences, ThermoPlex automatically designs multiplex PCR primer candidates in just a matter of minutes. The software also offers tools for thermodynamic calculations that can either be used apart from the automated primer screening routine or in conjunction with other existing primer design tools, depending on the on the needs of the user. Evidences presented here in this study provide insights on the performance of the software through theoretical and experimental analyses, serving to establish the reliability of its framework.

Keywords: DNA Thermodynamics; Multiplex PCR; Primer Design; Target-specific PCR

1 Introduction

Multiplex PCR-based assays remain widely employed for rapid DNA-based target detection due to its multi-target detection capabilities in a single assay platform [1]. Through the years, various detection and identification assays have emerged, utilizing this platform for a wide variety of organisms, including microbial [2, 3], plant [4], and animal [5, 6, 7, 8, 9] species. Recently, with the emergence of the coronavirus disease 2019 (COVID-19), numerous multiplex PCR assays for clinical diagnostics have been developed commercially and publicly [10, 11, 12, 13] taking advantage of the assay's ability to simultaneously detect the target along with multiple non-target samples in a single PCR reaction system. Multiplex PCR assays are thus expected to remain a relevant assay platform in the subsequent years, owing to their versatility in a variety of fields of application, as well as their cost-effectiveness.

The success of a PCR assay heavily depends on the design of its primers, perhaps its most vital component. In fact, PCR assay design issues typically arise from the designer's lack of experience with the primer design process [14]. This includes understanding of key parameters to optimize and familiarity with appropriate design tools to generate optimal primers. The complexity of the problem further escalates when trying to design a multiplex PCR assay. Apart from the primer's interaction with the target DNA, one must consider the interaction of multiple primers among themselves as all these components interact together in a single PCR solution system. These interactions result to numerous undesirable hybridization reactions affecting the assay's overall specificity and efficiency. Tackling such complexity with rigor, researchers often resort to an iterative, expensive and labor-intensive workflow [15] with high incidence of failure even after extensive experimentation [16].

The design challenges described can be alleviated by in silico approaches that systematically evaluate the performance of primers before undergoing experimental scrutiny. The most common type of in silico method for primer design evaluate primer performance through heuristics driven by practical experience with PCR (e.g., melting point matching, base pair mismatches, BLAST/alignment scores). However some of the major caveats of these approaches include the lack of direct physical interpretation, and reliance on parameters with arbitrarily chosen thresholds [17]. Such approaches can thus be considered lacking in terms of proper design of multiplex PCR assays. In contrast, approaches based on nucleic acid thermodynamics can overcome these limitations. Thermodynamics of nucleic acid interactions have been extensively studied in the past two decades, with values for useful thermodynamic parameters (enthalpy (ΔH), entropy (ΔS), and Gibbs free energy (ΔG) changes) already been compiled for every possible interaction motifs

[18, 19, 20, 21, 22, 23]. This makes it possible to quantitatively predict thermodynamic interactions between nucleic acids of any base pair sequence. Moreover, the use of thermodynamic parameters through the Nearest-Neighbor (NN) [24, 25] model have been proven to yield fairly accurate results consistent to that of experimental observations for oligonucleotide interactions [18]. This makes nucleic acid thermodynamics a solid foundation for developing a rigorous approach to designing multiplex PCR primers.

In this study, a design tool called ThermoPlex is introduced, with the aim of assisting with the design and screening of multiplex-compatible, target-specific PCR Primers from target and non-target sequence alignment datasets. Although numerous programs already exist predicting thermodynamic interactions of nucleic acids, very few open-access software focuses on automated and streamlined screening of multiplex PCR primer candidates from a given sequence dataset. The methodology utilizes a novel $O(M \times N)$ time algorithm based on the NN model doublet parameters to predict the ensemble thermodynamics of DNA-DNA interactions. The thermodynamic information is then used to negatively select target-specific primer candidates against undesirable target sequences. The target-specific primer candidates are then grouped together based on binding region overlaps and then assessed for multiplex-compatibility through an algorithm that computationally simulates multi-reaction equilibrium thermodynamics. The ThermoPlex software was designed in MATLAB with user interaction available either through a standalone GUI-enabled app (Windows/MacOS) or through the use of the source code. The standalone app together with the source codes are available at https://github.com/aagmata/ThermoPlex.

2 Methodology

2.1 *ThermoDHyb*: Thermodynamics of DNA hybridization prediction algorithm

The *ThermoDHyb* algorithm aims to predict the thermodynamics of double-stranded ensemble interaction between two single-stranded DNA. The novel algorithm calculates and predicts the associated Gibbs free energy change (ΔG) between a single-stranded reference state and each double-stranded microstate corresponding to local minima in the cumulative pairwise energy matrix P. The routine derives its main working principle from the Smith-Waterman algorithm [26] for sequence alignment. Sequence pairwise comparison (Figure 2A) is done per sequence doublet according to the NN model, with values of the P matrix populated according to the equation:

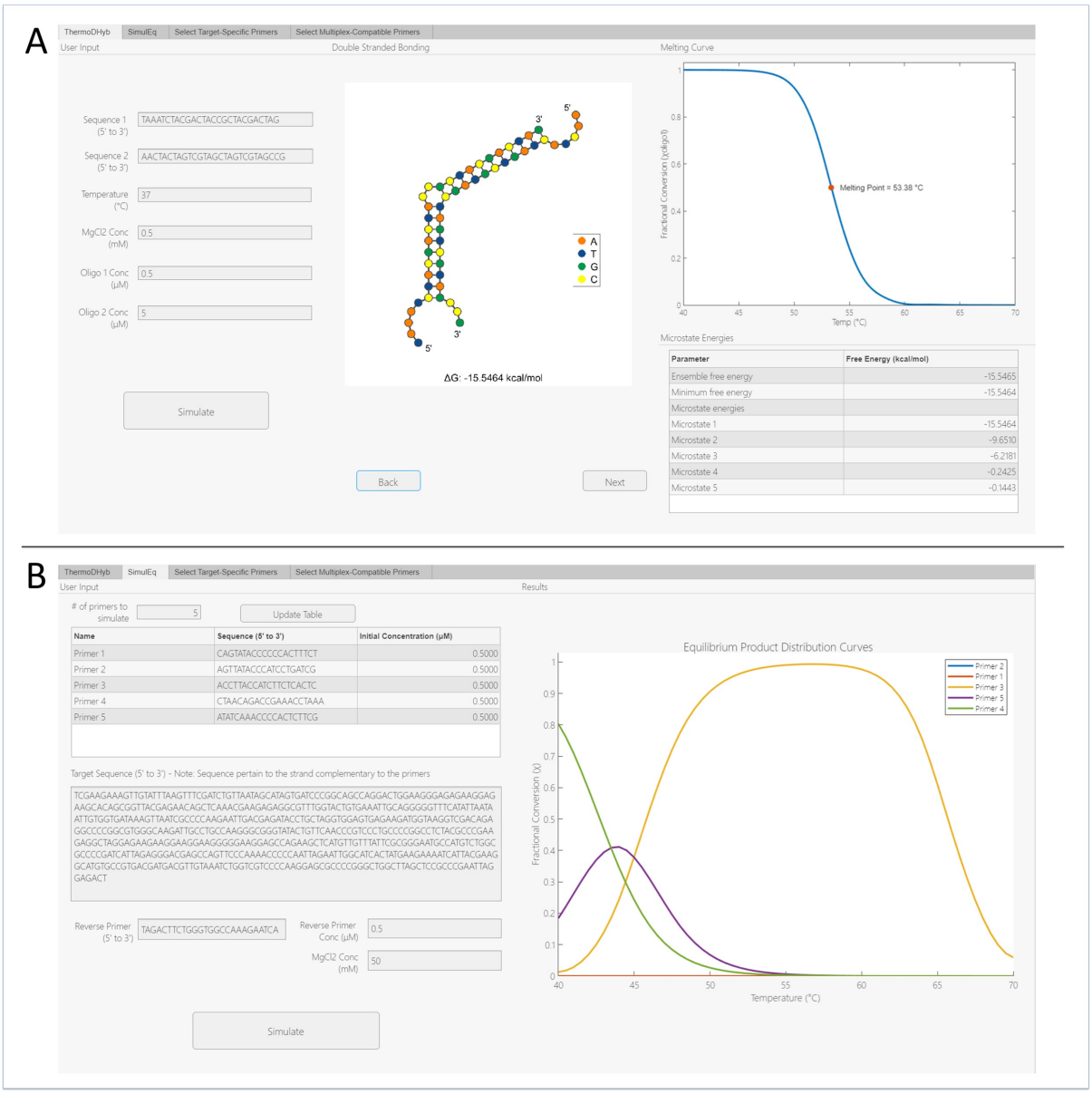

Figure 1: Snapshot of the ThemoPlex software user interface (UI) running the A) ThermoDHyb and B) SimulEq routines.

$$\Delta G_{m,n} = \min \begin{cases} \Delta G_{bu(m,n)} + \Delta G_{m-bu,n-1} + \Delta G_{\text{bulge}(1:bu)}, \\ \Delta G_{bl(m,n)} + \Delta G_{m-1,n-bl} + \Delta G_{\text{bulge}(1:bl)}, \\ \Delta G_{\text{NN}(m,n)} + \Delta G_{m-1,n-1} + \Delta G_{\text{loop}}, \\ 0 \end{cases} \tag{1}$$

where ΔG_{NN} is the cumulative free energy value, while ΔG_{bulge} and ΔG_{loop} correspond to length dependent penalties from bulge and internal loop formation, respectively. ΔG_{NN} values were calculated from enthalpy and entropy values while also accounting for the salt concentration dependence according to Owczarzy's model [27].

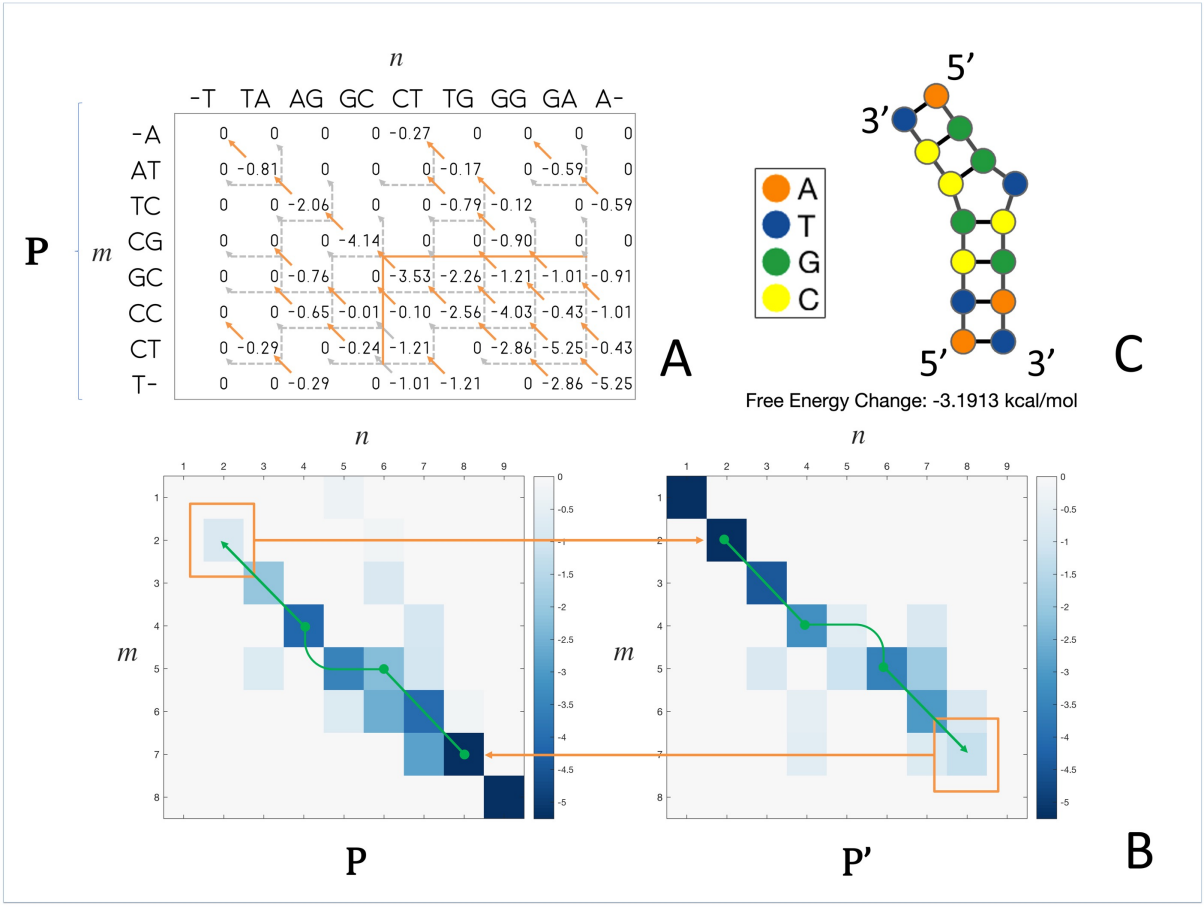

Figure 2: A) Cumulative pairwise energy matrix of doublet pairs for sequences 5' ATCGCCT 3' (Sequence i) and 5' AGGTCGAT 3' (Sequence j) at 40°C and 1M NaCl simulation conditions. Sequence i and j doublets are listed per row and column respectively. Arrows represent the cumulative dependence of each element $(m,\ n)$ to $(m - \text{bulge},\ n-1)$, $(m-1, n - \text{bulge})$, or $(m-1, n-1)$ elements. Orange, solid arrows depict the dependence case chosen corresponding to the minimum score case evaluated through equation 1. B) Propagation of cumulative energy scores (down-right for P, up-left for P'). Terminal doublet pairs are boxed in orange, green arrows show the path of tracing the doublets to infer the minimum free energy (MFE) structure of each microstate. C) Global MFE structure and corresponding free energy change $\Delta G^{i,j}_{\text{hyb}}$ of the most stable microstate predicted by *ThermoDHyb* for sequences i and j at 40°C and 1M NaCl simulation conditions.) *ThermoDHyb* and B) SimulEq routines.

The matrix-filling procedure is performed bi-directionally (Figure 2B) to predict multiple initiation and propagation paths across each local minimum, corresponding to each system microstate s. The energies of the set of microstates σ are then summed according to the partition function $Z_{i,j}$ (eq 2) which is in turn used to find the hybridization Gibbs free energy ($\Delta G^{i,j}_{\text{hyb}}$) according to equation 3:

$$Z_{i,j} = \sum_{s \in \sigma} e^{\frac{\Delta G_s^{i,j}}{RT}} \tag{2}$$

$$\Delta G^{i,j}_{\text{hyb}} = RT \ln(Z_{i,j}) \tag{3}$$

This allows *ThermoDHyb* to calculate thermodynamic interactions that account for significant contributors to the partition function weighed according to the Gibbs factor. Subsequently, the routine traces back the propagation path from the terminal state that along each local minimum to predict the minimum free energy (MFE) structure of each microstate (Figure 2C). The inner workings of the algorithm, together with assumptions made, are explained further in the Supplementary Material – Section 1.

2.2 *SiMulEq*: Simulation of Multi-reaction Equilibria interaction

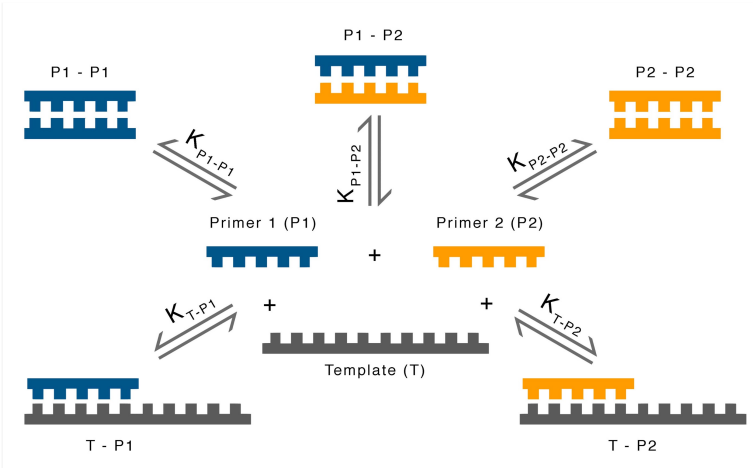

Figure 3: Complex multi-reaction equilibrium scenario accounting for each of the possible pairwise interactions in a two-primer, one-template PCR reaction system. Template-template interactions are not accounted for as it occurs on a different mechanism in the PCR annealing step.

The *SiMulEq* algorithm simulates the concentration of the double-stranded DNA products in thermodynamic equilibrium formed by the primers binding specifically and non-specifically to the given DNA template. The equilibrium scenario is modeled as a set of interdependent competitive reactions that seeks to minimize the total Gibbs free energy of the system. A fully-defined system of equation was derived from the minimization of the total Gibbs free energy as a function of all the species' chemical potential through the summability rule of partial properties [28]. The function was then constrained with material balance equations representing nucleotide strand conservation. By employing the method of Lagrange multipliers [29], the problem is converted from an optimization problem to a multi-dimensional root-finding problem of the form (derivation at Supplementary Material – Section 2):

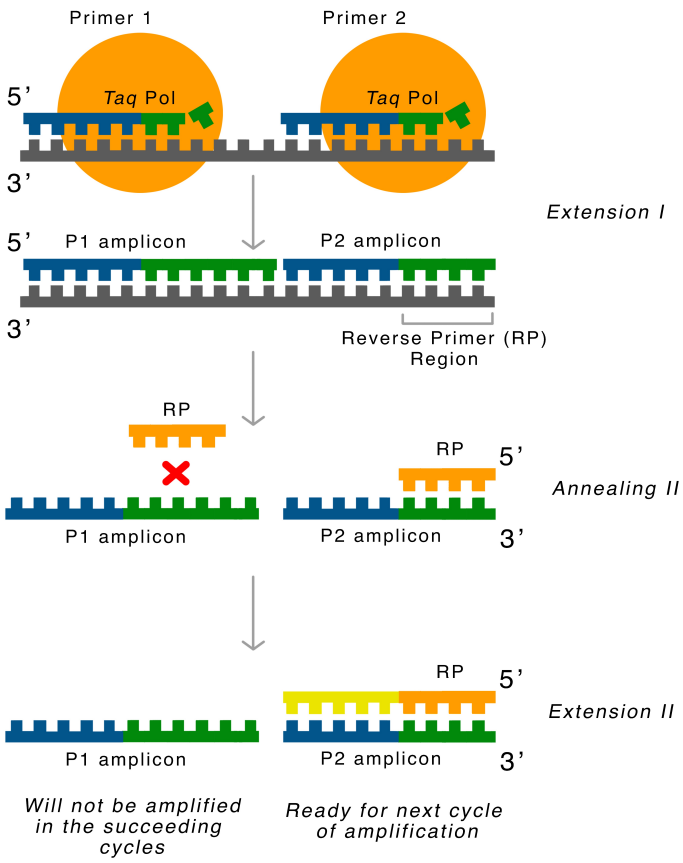

Figure 4: Extension inhibition mechanism where amplification of P1 products is inhibited by a primer (P2) binding along its extension path.

$$c_{0,i} - c_i - c_k^{ii} - \sum_k c_k^{ij} = 0 \quad (i = 1, 2, \ldots, p) \tag{4a}$$

$$\ln\left(\frac{c_i}{c_{0,i}}\right) + \sum_i \frac{\alpha_i \lambda_i}{RT} = 0 \quad (i = 1, 2, \ldots, p) \tag{4b}$$

$$\frac{\Delta G_{\text{hyb}}^{i,j}}{RT} + \ln\left(\frac{c_k^{ij}}{c_{0,i} c_{0,j}}\right) + \sum_i \frac{\alpha_i \lambda_i}{RT} = 0 \quad (i = 1, 2, \ldots, q) \tag{4c}$$

where $c_{0,i}$, c_i, c_k^{ii} are initial single-stranded, equilibrium single- and double-stranded DNA concenrations in the multiplex solution in standard mol/L, respectively, α_i is a repetition multiplier, λ_i is the lagrange multiplier for each strand species i and j. The resulting system of equations with $2p + q$ equations and unknowns are solved through Newton's method with finite-difference Jacobian. This yields concentration values of each chemical species at chemical equilibrium for any given temperature. The concentration values are then corrected (see Supplementary Material – Section 3) based on amplification inhibition

due to non-target primers binding between the forward target primer and the reverse primer (Figure 4). *SiMulEq* then iterates this procedure for a range of temperatures (40 – 70 °C) with 2 °C increments. The final output of *SiMulEq* are equilibrium product distribution (EPD) curves as functions of temperature.

2.3 *ThermoPlex* primer selection routine

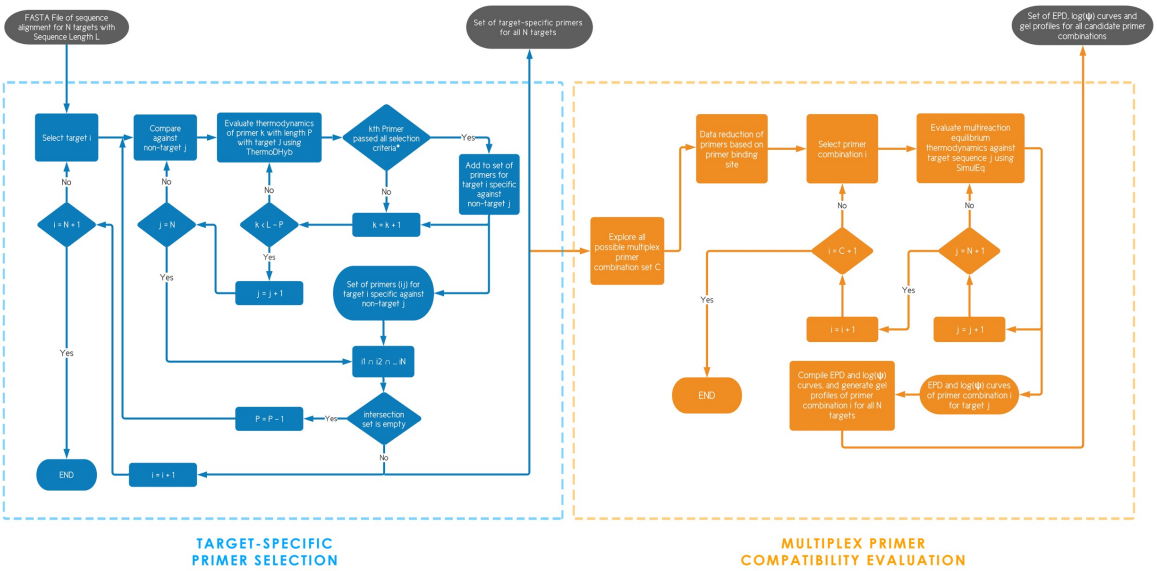

Figure 5: The *ThermoPlex* algorithm subdivided into the target-specific primer selection and multiplex primer compatibility evaluation. *Selection criteria include: no. of mismatch $<$ user-defined mismatch parameter (M); primer k does not contain ambiguous nucleotide characters; primer k does not form stable secondary structures (checked through MATLAB's bioinformatics toolbox function rnafold); fractional specificity (ψ) and target fractional conversion ($\chi_t|\Delta G^{t,p}_{\text{hyb}}$) $\geq$ user-defined threshold.

The ThermoPlex primer selection routine (Figure 5) aims to select multiplex-compatible, target-specific PCR primers from an input alignment of target sequences. Target-specific primers are first selected by iteratively evaluating every possible primer sequence from the consensus target sequence against all the non-target sequence through a moving frame with size equal to the primer length + 2 (the value 2 corresponds to single nucleotide flanking region at 5' and 3' end). This assumes that the most favorable interaction with a non-target template of a primer is located on the same region where the sequence was derived from the target sequence. In each iteration, a primer is first evaluated across a set of heuristic criteria (Figure 5*) prior to thermodynamic calculations. Subsequently, calculation of the primers $\Delta G^{i,j}_{\text{hyb}}$ on a single user-specified temperature (arbitrarily around PCR annealing temperature) with the sequence frame of non-target sequence j is performed through *ThermoDHyb*. Template fractional conversion χ_t, the mole fraction of the target or non-target template hybridized by the primer, is then calculated according

to the equation:

$$\chi_t = \frac{1}{2}\left[1+\alpha+\frac{\alpha}{c_{0,p}}\exp\left(\frac{\Delta G^{i,j}_{\text{hyb}}}{RT}\right)-\sqrt{\left(1+\alpha+\frac{\alpha}{c_{0,p}}\exp\left(\frac{\Delta G^{i,j}_{\text{hyb}}}{RT}\right)\right)^2-4\alpha}\right] \quad (5)$$

where:

$$\alpha = \frac{c_{0,p}}{c_{0,t}} \quad (6)$$

where $c_{0,t}$ and $c_{0,p}$, are the initial concentrations of the target template and the primer, respectively. Derivation is explained further in the Supplementary Material (Section 4). The derived model contains the variable $c_{0,t}$ in α, which is a huge source of uncertainty due to the difficulty of knowing its exact value in terms of molar units. To ensure the robustness of the model despite this uncertainty, a sensitivity analysis for all the relevant variables (T, MgCl2 concentration, and α) was performed to check whether the uncertainties in α is relevant enough. This was implemented through a Monte-Carlo simulation with uniform distribution of input variables.

Fractional specificity (ψ) is then calculated as the ratio of χ_t between target (T) to non-target (N):

$$\psi \equiv \frac{\chi_t\big|_{\Delta G^{T,p}_{\text{hyb}}}}{\chi_t\big|_{\Delta G^{N,p}_{\text{hyb}}}} \quad (7)$$

ψ describes how probable a primer will hybridize with the target compared to the undesirable non-target sequence (e.g. $\psi = 1000$ means there is 1000 times more primer-target product compared to primer-non-target product upon annealing at the specified temperature). Both ψ and $\chi_t \mid \Delta G^{T,p}_{hyb}$ are used as selection criteria based on a user-defined threshold to screen for primer specificity and efficiency, respectively. Primers that passed the criteria are then included in the set ij corresponding to primers derived from target i specific against non-target j. Upon completion of iteration through all the non-target sequences, the intersection of sets i_1, i_2, ... i_N (where N is the total number of targets) is then taken, comprising the set of primers for target i specific to all the non-target sequences. In the event that the intersection results to an empty set, the process is repeated for a reduced primer length to increase the relative effects of mismatches and thus

decrease $\chi_t \mid \Delta G_{hyb}^{N,p}$, increasing primer specificity. The whole process is repeated until each target has their own set of target-specific primers.

The set of all target-specific primers for each target are then used to build all possible primer combinations with one primer each target, for a total of N primers per combination. The combinations are evaluated on the constraint of amplicon site resolution (R). Specifically, each primer in a multiplex combination should be spaced at least R bp apart. This will ensure that primers will produce distinguishable gel bands corresponding to each of the target in a multiplex set. Each primer combinations are then evaluated on their multi-reaction equilibrium thermodynamics using *SimulEq*. New values for fractional specificity are also calculated in the context of the multiplex reaction system ($\psi_i(T)$) for each primer i corresponding to the desired highest yielding primer, across each temperature increment T:

$$\psi_i(T) = \frac{c_{i,t}}{\sum_{j \neq i} c_{j,t}} \tag{8}$$

This results to a curve describing the fractional conversion of target t from its hybridization with primer i relative to all the undesired double-stranded products l due to hybridization between primers j and target t. The routine will ultimately output EPD curves, $log(\psi)$ curves and gel profiles for each multiplex primer combination which gives insights on the performance of each multiplex primer set across varying temperature conditions. The ThermoPlex's subroutines ("Select Target-Specific Primers" and "Select Multiplex-Compatible Primers") comprising the whole primer selection process were benchmarked in terms of runtimes, with input parameters listed in Table 1, to test how fast each performs across various personal computing platforms.

Table 1: ThermoPlex parameters used in the study for the design of Multiplex PCR Primers

Parameters	**Defined or set as threshold**
$T_{\text{simulation}}$	55°C
$MgCl_2$ concentration	10 mmol/L
Equimolar primer concentration ($c_{0,p}$)	0.5 μmol/L
Frac Specificity at $T_{\text{simulation}}$ ($\psi \mid T_{\text{simulation}}$)	1000
Frac Threshold at $T_{\text{simulation}}$ ($\chi_t \mid \Delta G_{t,phyb}$)	0.5
Initial primer length (P)	20
Mismatch parameter (M)	3
Amplicon length resolution (R)	50

2.4 Laboratory Validation

Performance of the predictive capabilities of the *SimulEq* algorithm in simulating multi-reaction equilibrium primer interactions was tested using an in-house multiplex assay for identifying five (5) sardine species through the amplification of Cytochrome C Oxidase I (COI) gene. A simulation was performed using the primer sequences listed in Table 2 and then compared with gel electrophoresis profile of the Assay. Meanwhile, the capabilities of the *ThemoPlex* algorithm to assist in the screening of target-specific multiplex PCR primers were also assessed. Multiplex PCR primers (Table 2), also targeting the COI gene region, were designed for an assay to delineate four (4) sardine species using the algorithm. One of the three predicted set of primers were then randomly picked to be evaluated in the laboratory for its performance.

Table 2: Target species sample set for the PCR trial run with corresponding specific primers and expected amplicon product size for both In-house and ThermoPlex-designed primers.

Target Species	Sequence (5' to 3')	Approx. product size (bp)
	In-house Primers	
S.fimbriata	CTAACAGACCGAAACCTAAA	121
S.goni	-	-
S.lemuru	ATATCAAACCCCACTCTTCG	215
H.quadrimaculatus	AGTTATACCCATCCTGATCG	542
S.tawilis	ACCTTACCATCTTCTCACTC	318
A.sirm	CAGTATACCCCCCACTTTCT	369
	Designed Primers	
S.fimbriata	TATTACTACGATTATCAACA	121
S.goni	CTCCGTCGACCTAACTATTT	239
S.lemuru	GATCAAATCTACAATGTTAT	511
H.quadrimaculatus	GAAATTTAAACACAACCTTC	24
S.tawilis	-	-
A.sirm	-	-

High-quality DNA used for the experiment were extracted from tissue samples of each of the target species using Qiagen's DNeasy Blood and Tissue Kit. The extracted DNA was amplified through PCR using a 10 μL PCR mix comprised of 2.45 μL of ultrapure water, 2 μL of 5x GoTaq® PCR Buffer, 0.4 μL of 25mM $MgCl_2$, 0.8 μL of 2.5mM dNTP, 5 x 0.5 μL of 10 µM each of the five forward primers, 0.5 μL of 10 μM reverse primer, 0.8 μL of Bovine Serum Albumin (BSA), 0.05 μL of 5 u/μL GoTaq® polymerase and 0.5 μL of DNA template. Gel electrophoresis was then performed using 3% agarose gel to view the results of the amplification.

3 Results and Discussion

3.1 Benchmarking of *ThermoDHyb* algorithm and *ThermoPlex* subroutines

Table 3: Resulting fitting parameters for each model through non-linear regression of ThermoDHyb runtimes

	Fitting Parameter	
Model	**a**	**b**
$ax + b$	1.85 $\times 10^{-2}$	-0.4037
$ax^2 + b$	1.59 $\times 10^{-2}$	0.0326
$ax^3 + b$	1.57 $\times 10^{-6}$	0.1930
$ax^4 + b$	1.57 $\times 10^{-8}$	0.2824

Table 4: Computing platform specifications, subroutine average runtimes (3 replicates), and input parameters for the ThermoPlex benchmarking

	1	**2**	**3**
Specifications			
Model	ASUS TUF A15	MSI GL63 8RD	MacBook Pro 2019
Operating System	Windows 10	Windows 10	MacOS Catalina (10.15.7)
Processor	AMD Ryzen 7 4800H	Intel® Core™ i7-8750	Intel® Core™ i5-8279U
No. of Cores	8	6	4
Clock rate	2.9 GHz	2.20 GHz	1.4 GHz
RAM	8 GB	12 GB	8 GB
Average runtimes			
Select target-specific primers	54 s	1 min 32 s	1 min 44 s
Select multiplex-compatible primers	3 min 43 s	5 min 5 s	5 min 12 s
TOTAL	4 min 37 s	6 min 37 s	6 min 56 s
Benchmark parameters			
No. of target sequences		4	
Sequence alignment length		563	
No. of predicted multiplex combinations		3	

Table 3 provides the list of model equations used to assess the computational complexity of the *ThermoDHyb* algorithm, and their corresponding fitted parameters. Visual inspection of the results of the time-complexity benchmark (Figure 6) suggests that the quadratic model is the best fit for the resulting simulation runtimes as it demonstrates good overlap with the data points relative to other polynomial models. Moreover, the b fitting parameter (y-intercept) of the quadratic model has the closest value to zero suggesting the model's ideal fit compared to others. All these indicates an average-case quadratic time complexity ($O(N^2)$) for the ThermoDHyb algorithm, for the special case of equal sequence length ($M = N$). The general case is thus expected to be $O(M \times N)$

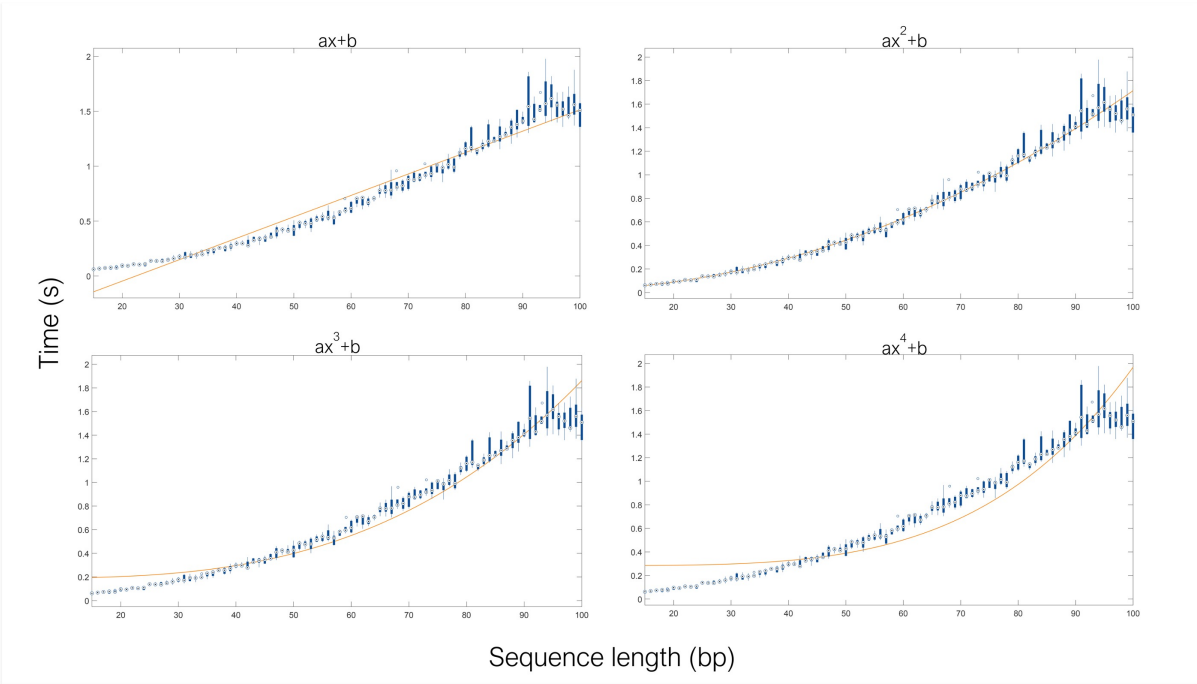

Figure 6: Benchmarking for time-complexity of *ThermoDHyb* through exploration of runtime as a function of sequence length (bp). Box plots (blue) were generated from sequence length of 15 to 100 bp at 1 bp increment with 5 replicates each bp. The orange lines represent each polynomial model fitted through non-linear least squares.

for $M \neq N$ as this represents the computational bottleneck of filling up the $m \times n$ matrix of doublet interactions. The sub-routine's efficiency better fits the iterative nature of the ThermoPlex algorithm compared to other published algorithms predicting nucleic acid thermodynamic interactions such as McCaskill's [30] and Dirks's [31] both running on $O(N^4)$ time. Although their algorithm's sampling of possible microstates in the ensemble is more rigorous, the microstates predicted in *ThermoDHyb* should be enough to sample the most important contributors in the ensemble free energy, particularly in the case of PCR applications where interactions are limited to small oligonucleotide primers that seldom adopt more complex secondary structures and interactions such as pseudoknots. Indeed, the benchmark results (Table 4) manifest the algorithm's efficiency, with subroutines completing the computations under 10 minutes across various personal computers and operating systems tested in the study.

Meanwhile, results of the linear regression analysis (Figure 7) between ThermoDHyb and three other software (Visual OMP [32], (B) DINAMelt [33], and (C) NUPACK [34]) suggest good concordance in the predicted ΔG values, especially with Visual OMP. A near-unity regression slope and R^2 of 0.9959 and 0.9770, respectively, as well as a near-zero y-intercept of 0.0192 indicates a good one-to-one agreement in the predicted values. Good concordance (84%) in terms of predicted secondary structures is also observed, with non-concordance existing only at higher ΔG values (> -3 kcal/mol). These sug-

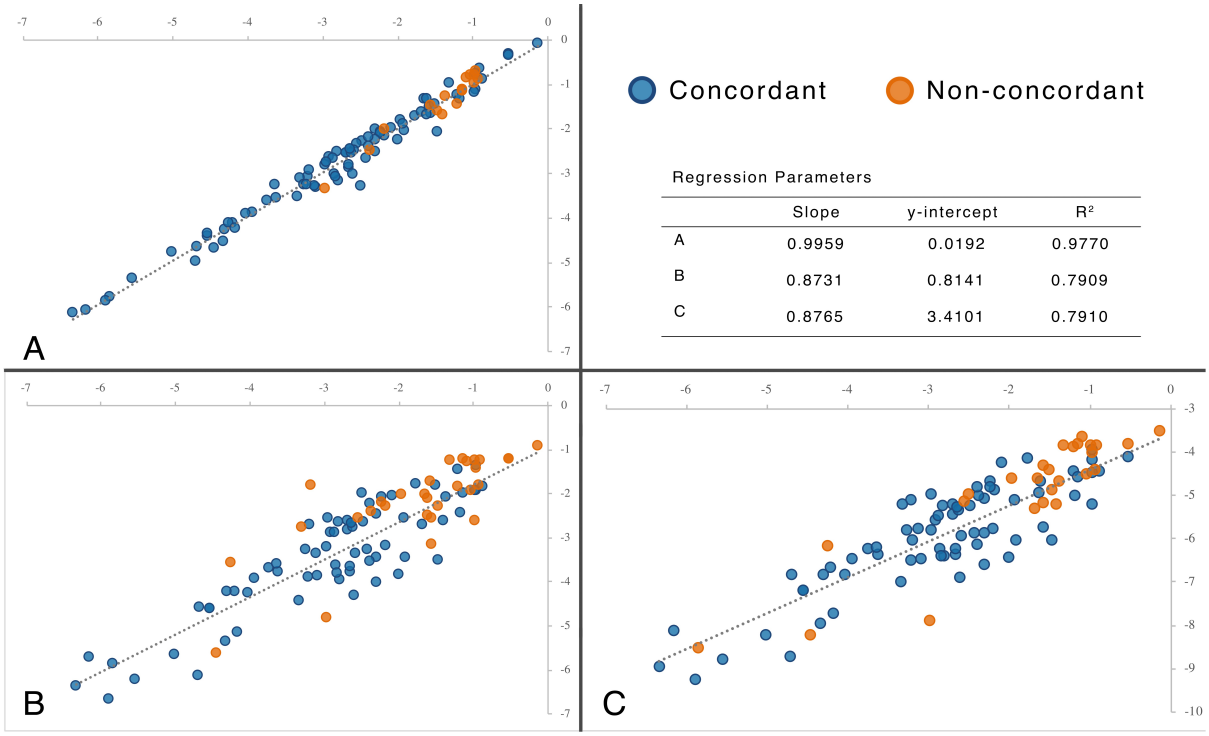

Figure 7: Linear regression analysis for the benchmarking of predicted global minimum free energy values (kcal/mol) simulated at 37°C and 1M NaCl solution as predicted by *ThermoDHyb* as compared to various software: (A) Visual OMP [32], (B) DINAMelt [33], and (C) NUPACK [34]. Blue and orange data points correspond to concordant and non-concordant predicted secondary structures between *ThermoDHyb* and the software.

gest that the *ThermoDHyb* algorithm is at least at par with the accuracy of the Visual OMP software in terms of predicting the thermodynamic interactions of two oligonucleotides. Although quite expensive, the Visual OMP software has been extensively used in numerous multiplex-PCR primer design studies [35, 36, 37, 38] and yielded accurate results. Moreover, fairly good concordance in predicted secondary structures (70% and 84%) and ΔG values, at least in terms of trend, can still be seen in the comparison with both DINAMelt and NUPACK. It is also worth noting that the y-intercept of Figure 7C implies a shift in the predicted value of 3.41 kcal/mol on average by NUPACK. This shift is consistent with the study of Zhang et al. [39] where they needed to adjust the NUPACK-predicted ΔG values by 1.5 kcal/mol for their model to better fit the experimental results. All of these results establish the reliability of the ThermoDHyb algorithm and thus ensuring a solid framework for thermodynamic calculations of oligonucleotide interactions performed within the ThermoPlex routines.

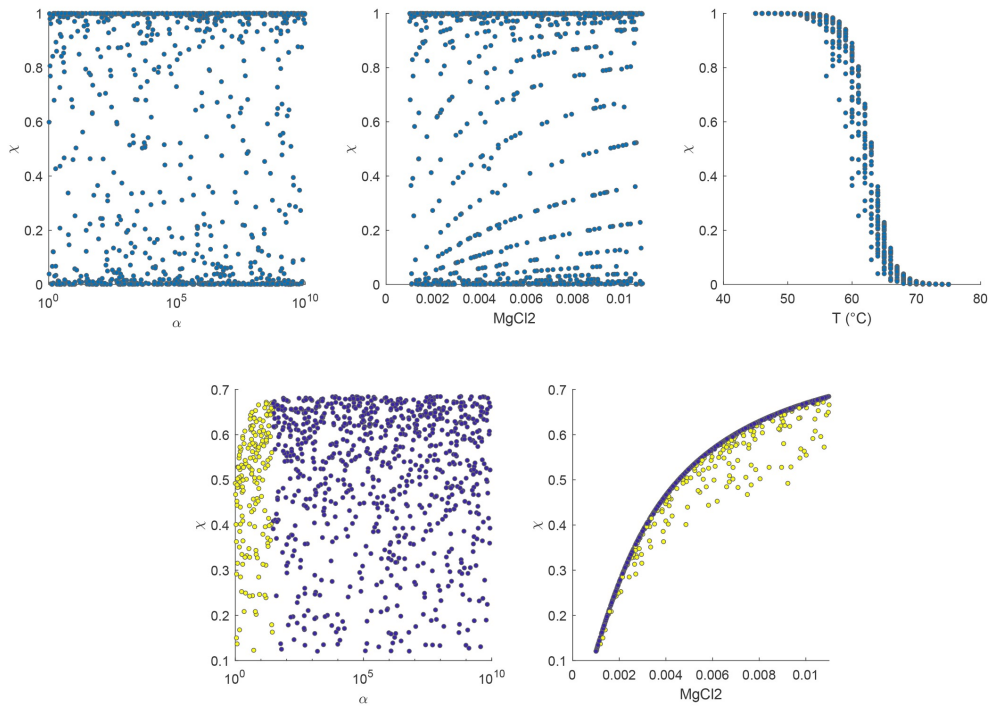

Figure 8: Monte Carlo (MC) simulation ($n = 1000$) scatter plots of χ versus α (A and D), $MgCl_2$ concentration (B and E), and T (C). The parameter T was held constant at 62°C for MC simulations ($n = 1000$) in D and E.

3.2 Sensitivity analysis of the χ model for fractional conversion

The scatter plots from Monte Carlo (MC) simulations in Figure 8A-C show that χ is most sensitive to temperature parameter changes, followed by $MgCl_2$ concentration, and lastly α. This is elucidated by a more distinct pattern in the plot of C than in B, and then in A. Such is an expected trend since DNA hybridization is most strongly influenced by temperature as heat is mainly responsible for breaking the hydrogen bonds of the DNA's N-bases in PCR reactions. Meanwhile, salt concentration moderately influences DNA hybridization stability through the neutralization of the negative charges on the backbones by cationic metals (Na^+ and Mg^{2+}), reducing electrostatic repulsion between phosphates [40]. On the other hand, no pattern is discernable in plot A, but further evaluation was performed since the pattern may be heavily diluted by high variations in T. This was done by performing another MC simulation where T is held constant at 62°C (Figure 8 D-E), thus comparing only the influence of $MgCl_2$ and α parameters. The results of the second simulation depict that although a pattern is completely discernable in plot E and still none in plot D, outliers (yellow) can be observed in the former indicating that α still influences the model to a certain extent. However, linking these outliers to plot D (also yellow) suggests that the significant influence of α is limited only to lower

limit values ($\alpha < 30$), while the influence diminishes as values of α goes higher (blue). PCR reactions are typically conducted with α of at least 108 to ensure that template will anneal to primers rather than itself [41], thus the model is sufficiently valid for PCR applications despite the uncertainties attributed to α. The simulation results also provide an interesting insight in the context of modelling any bimolecular interactions beyond PCR applications, as the χ function was derived from a generalized mass action equation for bimolecular reactions. The model can thus be used for predicting fractional conversion of a limiting reactant even without accurate information about its value, given that the excess reactant is at least at a factor of 102 more than the limiting.

3.3 *SimulEq* and *ThermoPlex* Simulations, and Laboratory Validation

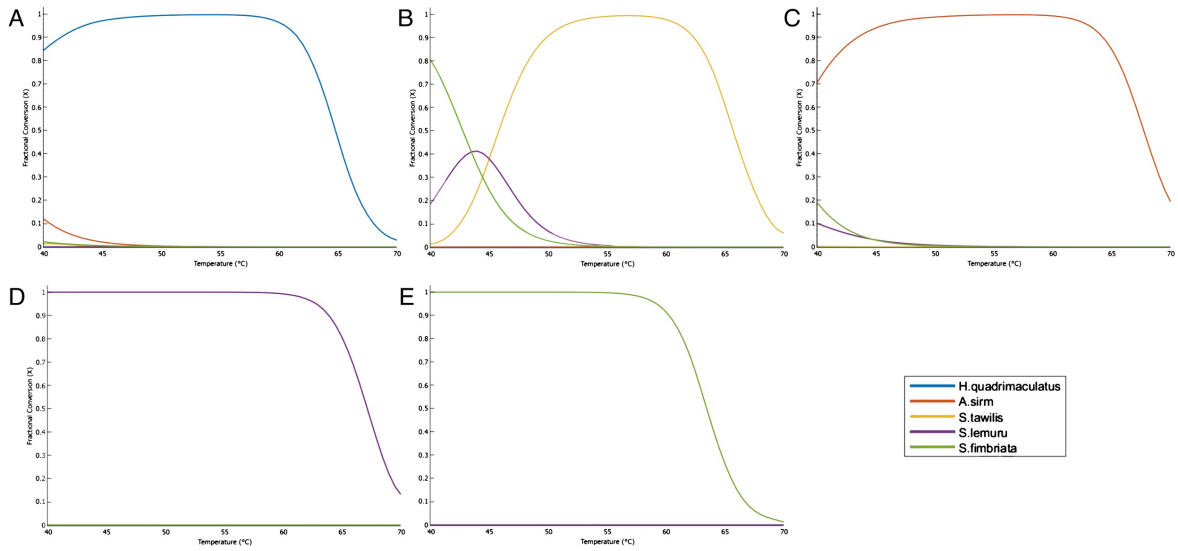

Figure 9: Equilibrium Product Distribution (EPD) curves as function of temperature generated by *SimulEq* from the in-house developed multiplex PCR assay primers simulated per target COI gene sequence. Each color corresponds to the target species the primer is designed for. Primer sequences and their targets are listed in Table 2.

Generated EPD curves from Figure 9 through the *SimulEq* algorithm illustrates the predicted product distributions of the whole multiplex assay as functions of temperature. Each plot corresponds to the multiplex assay's performance on each target species' gene sequence. Inspection of the resulting EPD curves suggest that the assay's primers for *S. lemuru* and *S. fimbriata* are completely specific to their own respective target species. This is because in plots D and E, apart from products from their respective targets, there are no other curves visible. Such is the opposite to plots A to C as multiple curves are visible aside from each of their own target curves, particularly the primer for *S. tawilis* (plot B) wherein multiple curves are already obvious for as a high a temperature as

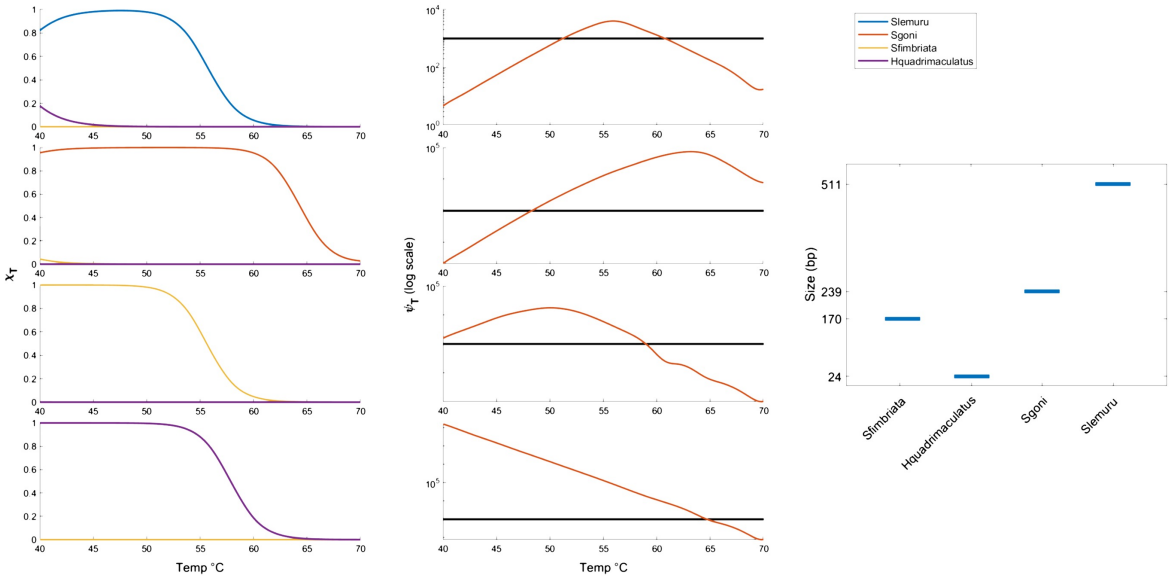

Figure 10: Output of the *ThermoPlex*'s design algorithm showing A) EPD curves, B) $\psi(T)$ curves (log10 y-axis scale) as functions of temperature (black lines correspond to $\psi = 1000$), and C) predicted gel electrophoresis profiles for the multiplex assay.

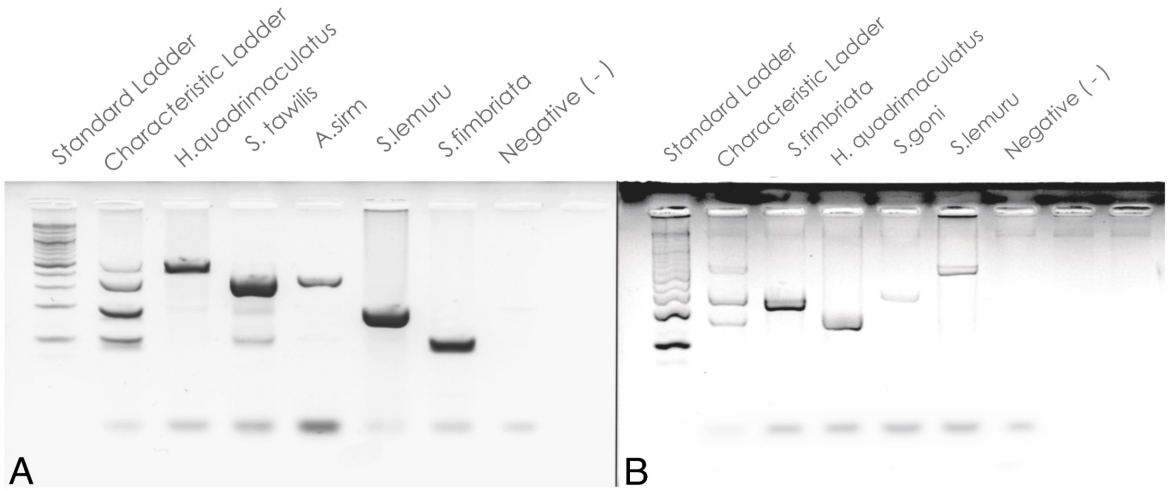

Figure 11: Gel electrophoresis profile of the PCR products for A) in-house developed multiplex PCR assay and B) *ThermoPlex*-designed the multiplex PCR assay.

55°C. This implies that the primer for *S. tawilis* should demonstrate a certain degree of non-specificity towards *S.lemuru* and *S. fimbriata*. Indeed, this is consistent with experimental results as shown in Figure 11A. Gel electrophoresis profile for the assay indicates the same expected non-specificity of *S. tawilis* primers for both species, as shown by the two faint gel bands in its lane whose sizes are consistent with that of *S. lemuru* and *S. fimbriata* PCR products. The consistency of *SimulEq* simulations with experimental results demonstrates the reliability of the algorithm in the thermodynamic analysis of multiplex PCR assays. It is worth noting, however, that this type of theoretical analysis

does not directly translate to expected PCR products per se due to the simulation capturing only what is happening per annealing cycle, or the first cycle at best. Nevertheless, the analysis done on the EPD curves still proves to be insightful since it illustrates the assay's behavior when hybridization due to non-specific primer binding becomes more and more significant at lower temperatures.

The output of the ThermoPlex design routine is shown in Figures 10 and 11 which includes the multiplex specificity ($\psi_i(T)$) curves and the predicted gel electrophoresis profile for the designed assay. Due to the unavailability of a scoring function (still under development) to rank all the designed candidate multiplex primer sets, random selection of one assay out of three candidate primer sets were done for laboratory validation. Following the same analysis of the EPD curves as above suggests that the primers in the designed assay are expected to be more target-specific than the in-house developed primers. The ψ_i curves in Figure 11B also aids in this inference as it illustrates the expected fractional specificity of the primer set at various temperatures. It shows the temperature limits at which each primer in the assay should theoretically retain its specificity above the user-defined ψ threshold (black horizontal line). Moreover, the curves suggest that at Tsimulation of 55°C, the assay retains specificity, with the least specific primer (*S. lemuru*) still having ψ_i values in the order of 104, i.e., the primer is expected to hybridize 104 times more to the target than all the other primers in the assay combined. The result of the simulation is indeed supported by the experimental results in Figure 11B. The gel electrophoresis profile of the designed multiplex assay shows no discernable non-specificity, with only one distinct gel band manifesting per each target. The experimental results thus show that the set of multiplex primers automatically designed by the *ThermoPlex* routine performs adequately, proving the reliability of the primer screening routine the software implements. Tests involving larger data sets (e.g. longer sequence length, more targets), however, should be performed to further assess the automated screening capabilities of *ThermoPlex*.

4 Conclusion and recommendations

Despite the proved performance of ThermoPlex in automatically designing multiplex PCR primers, one caveat is the limitation in terms of sequence input to the program, specifically:

1. The program strictly needs the sequence input to be aligned and trimmed to equal length.

2. In its current state, it is unable to handle sequence alignments with gaps, i.e., gene regions with insertions and deletions.

3. It also fails to pick candidates for sequences with highly dispersed and sparse polymorphic sites despite having fairly high genetic distances.

Moreover, the laboratory performance assessments so far of ThermoPlex and all of its subroutine lack sufficient quantitative basis despite the program being motivated by quantitative thermodynamic variables. Hence, it would be best to test its predictions with quantitative experiments such as qPCR-based methodologies. These types of experiments would also empirically fine-tune the models underlying the algorithm's framework, making them more consistent to real-life expectations.

Input-parameter optimization was also not done in order to exhaustively explore the output candidate primer space to design more optimal multiplex primers. Additionally, a ranking/scoring system would also be needed, especially if the program were able to select numerous candidate primer sets that will be difficult for the user to evaluate and handle. Presently, the workaround would be to make the parameters (e.g., $MgCl_2$ concentration, specificity thresholds) stringent enough to produce a manageable number of candidate sets. All of these will be implemented in the future to further optimize the selection process.

In summary, the ThermoPlex program demonstrates great potential to improve the accuracy and speed of designing multiplex PCR primers through all the evidence presented in this study. With the program's automated screening routine, designers can expect to have candidate multiplex PCR primers within minutes, subject to the speed of the computing platform, number of design targets, length of the input sequence, and number of candidate multiplex set predicted. All these are achievable through common personal computers without the need for high performance computing. Moreover, apart from the automated primer screening algorithm, the program also offers useful and insightful tools motivated by DNA thermodynamics via its ThermoDHyb and SimulEq routines. Depending on the needs of the user, these novel algorithms can be used independently or in concert with other existing programs. Although already useful in its current state, further improvements in the future are to be implemented to enhance the program's capabilities such as more robust handling of sequence alignment, ability to score and rank predicted multiplex sets, and possible expansion to a web server format.

5 Availability

The standalone app together with the source codes are available in the GitHub repository (https://github.com/aagmata/ThermoPlex)

6 Acknowledgement

The authors would like to acknowledge the contribution of Dr. Ma. Rio. Naguit, Dr. Asuncion De Guzman, Mr. Jerry Garcia, Mr. Jhunrey Follante, and Mr. John Christopher Azcarraga in the collection and processing of sardine samples. Additionally, the authors would like to express their gratitude to Dr. Dame Loveliness Apaga-Agmata for her valuable insights in chemical thermodynamics and physical chemistry.

7 Funding

The work was made possible through the funding support by the Department of Science and Technology – Philippine Council for Agriculture, Aquatic, and Natural Resources Research and Development (DOST-PCAARRD);

References

[1] Ali, M. E., Razzak, M. A., and Hamid, S. B. A. (11, 2014) Multiplex PCR in Species Authentication: Probability and Prospects—A Review. *Food Analytical Methods,* **7**, 1933–1949.

[2] Aguilera-Arreola, M. G., Gonzalez-Cardel, A. M., Tenorio, A. M., Curiel-Quesada, E., and Castro-Escarpulli, G. (2014) Highly specific and efficient primers for in-house multiplex PCR detection of Chlamydia trachomatis, Neisseria gonorrhoeae, Mycoplasma hominis and Ureaplasma urealyticum.. *BMC research notes,* **7**, 433.

[3] Ke, L., Wang, L., Li, H., Lin, H., and Zhao, L. (7, 2014) Molecular identification of lactic acid bacteria in Chinese rice wine using species-specific multiplex PCR. *European Food Research and Technology,* **239**, 59–65.

[4] Moon, J. C., Kim, J. H., and Jang, C. S. (2016) Development of multiplex PCR

for species-specific identification of the Poaceae family based on chloroplast gene, rpoC2. *Applied Biological Chemistry,* **59**, 201–207.

[5] Ali, M. E., Razzak, M. A., Hamid, S. B. A., Rahman, M. M., Amin, M. A., Rashid, N. R. A., and Asing (2015) Multiplex PCR assay for the detection of five meat species forbidden in Islamic foods. *Food Chemistry,* **177**, 214–224.

[6] Hou, B., Meng, X., Zhang, L., Guo, J., Li, S., and Jin, H. (2015) Development of a sensitive and specific multiplex PCR method for the simultaneous detection of chicken, duck and goose DNA in meat products. *Meat Science,* **101**, 90–94.

[7] Piergiorge, R. M., Pontes, M. N., Duarte, A. V. B., and Gusmão, J. (2014) Haplotype-specific single-locus multiplex PCR assay for molecular identification of sea-bob shrimp, Xiphopenaeus kroyeri (Heller, 1862), cryptic species from the Southwest Atlantic using a DNA pooling strategy for simultaneous identification of multiple s. *Biochemical Systematics and Ecology,* **54**, 348–353.

[8] You, J., Huang, L., Zhuang, J., and Mou, Z. (2014) Species-specific multiplex real-time PCR assay for identification of deer and common domestic animals. *Food Science and Biotechnology,* **23**, 133–139.

[9] Ravago-Gotanco, R. G., Manglicmot, M. T., and Pante, M. J. R. (2010) Multiplex PCR and RFLP approaches for identification of rabbitfish (Siganus) species using mitochondrial gene regions. *Molecular Ecology Resources,* **10**, 741–743.

[10] Ishige, T., Murata, S., Taniguchi, T., Miyabe, A., Kitamura, K., Kawasaki, K., Nishimura, M., Igari, H., and Matsushita, K. (8, 2020) Highly sensitive detection of SARS-CoV-2 RNA by multiplex rRT-PCR for molecular diagnosis of COVID-19 by clinical laboratories. *Clinica Chimica Acta,* **507**, 139–142.

[11] Mancini, F., Barbanti, F., Scaturro, M., Fontana, S., Martino, A. D., Marsili, G., Puzelli, S., Calzoletti, L., Facchini, M., Mario, G. D., Fabiani, C., Bella, A., Riccardo, F., Pezzotti, P., Stefanelli, P., Rezza, G., Ciervo, A., Villa, L., Fortini, D., Iacobino, A., Fiore, S., Benedetti, E., Marchi, A., Venturi, G., Fortuna, C., Amendola, A., Toma, L., Luca, M. D., and Severini, F. (10, 2020) Multiplex Real-Time Reverse-Transcription Polymerase Chain Reaction Assays for Diagnostic Testing of Severe Acute Respiratory Syndrome Coronavirus 2 and Seasonal Influenza Viruses: A Challenge of the Phase 3 Pandemic Setting. *The Journal of Infectious Diseases,*.

[12] Park, M., Won, J., Choi, B. Y., and Lee, C. J. (2020) Optimization of primer sets and detection protocols for SARS-CoV-2 of coronavirus disease 2019 (COVID-19) using PCR and real-time PCR. *Experimental and Molecular Medicine,* **52**, 963–977.

[13] Kudo, E., Israelow, B., Vogels, C. B. F., Lu, P., Wyllie, A. L., Tokuyama, M., Venkataraman, A., Brackney, D. E., Ott, I. M., Petrone, M. E., Earnest, R., Lapidus, S., Muenker, M. C., Moore, A. J., Casanovas-Massana, A., Omer, S. B., Cruz, C. S. D., Farhadian, S. F., Ko, A. I., Grubaugh, N. D., and Iwasaki, A. (10, 2020) Detection of SARS-CoV-2 RNA by multiplex RT-qPCR. *PLOS Biology,* **18**, e3000867.

[14] Bustin, S. and Huggett, J. (12, 2017) qPCR primer design revisited. *Biomolecular Detection and Quantification,* **14**, 19–28.

[15] Henegariu, O. and Heerema, N. (1997) Multiplex PCR: critical parameters and step-by-step protocol. *BioTechniques,* **23**, 504–11.

[16] Yuryev, A. (2007) PCR Primer Design, Vol. 402, Humana Press, .

[17] Mann, T., Humbert, R., Dorschner, M., Stamatoyannopoulos, J., and Noble, W. S. (2009) A thermodynamic approach to PCR primer design. *Nucleic Acids Research,* **37**.

[18] SantaLucia, J. and Hicks, D. (2004) The thermodynamics of DNA structural motifs.. *Annual review of biophysics and biomolecular structure,* **33**, 415–40.

[19] Bommarito, S., Peyret, N., and SantaLucia, J. J. (2000) Thermodynamic parameters for DNA sequences with dangling ends. *Nucleic acids research,* **28**, 1929–1934.

[20] SantaLucia, J. (1998) A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics.. *Proceedings of the National Academy of Sciences of the United States of America,* **95**, 1460–1465.

[21] Allawi, H. T. and SantaLucia, J. (1998) Thermodynamics of internal C??T mismatches in DNA. *Nucleic Acids Research,* **26**, 2694–2701.

[22] Allawi, H. T. and SantaLucia, J. (1998) Nearest neighbor thermodynamic parameters for internal G??A mismatches in DNA. *Biochemistry,* **37**, 2170–2179.

[23] Allawi, H. T. and Jr., J. S. (1997) Thermodynamics and NMR of internal G.T mismatches in DNA. *Biochemistry,* **36**, 10581–10594.

[24] Crothers, D. M. and Zimm, B. H. (1964) Theory of the melting transition of synthetic polynucleotides: Evaluation of the stacking free energy. *Journal of Molecular Biology,* **9**, 1–9.

[25] Kallenback, N. R. and Crothers, D. M. (1966) Theory of thermal transitions in cohered DNA from phage lambda. *Proc Natl Acad Sci U S A,* **56**, 1018–1025.

[26] Smith, T. F. and Waterman, M. S. (1981) Identification of common molecular subsequences.. *Journal of molecular biology,* **147**, 195–7.

[27] Owczarzy, R., Moreira, B. G., You, Y., a. Behlke, M., and a. Wälder, J. (2008) Predicting stability of DNA duplexes in solutions containing magnesium and monovalent cations. *Biochemistry,* **47**, 5336–5353.

[28] Dimitrov, R. and Zuker, M. (2004) Prediction of hybridization and melting for double-stranded nucleic acids.. *Biophysical journal,* **87**, 215–26.

[29] Smith, J. M.; Van Ness, H. C. A. (2005) Introduction to Chemical Engineering Thermodynamics, McGraw-Hill, 7th edition.

[30] McCaskill, J. S. (1990) The equilibrium partition function and base pair binding probabilities for RNA secondary structure.. *Biopolymers,* **29**, 1105–1119.

[31] Dirks, R. M., Bois, J. S., Schaeffer, J. M., Winfree, E., and a. Pierce, N. (2007) Thermodynamic Analysis of Interacting Nucleic Acid Strands. *SIAM Review,* **49**, 65–88.

[32] SantaLucia, J. Physical Principles and Visual-OMP Software for Optimal PCR Design pp. 3–33 (2007).

[33] Markham, N. R. and Zuker, M. (7, 2005) DINAMelt web server for nucleic acid melting prediction. *Nucleic Acids Research,* **33**, W577–W581.

[34] Zadeh, J. N., Steenberg, C. D., Bois, J. S., Wolfe, B. R., Pierce, M. B., Khan, A. R., Dirks, R. M., and Pierce, N. A. (1, 2011) NUPACK: Analysis and design of nucleic acid systems. *Journal of Computational Chemistry,* **32**, 170–173.

[35] Medina, R. A., Rojas, M., Tuin, A., Huff, S., Ferres, M., Martinez-Valdebenito, C., Godoy, P., Garcia-Sastre, A., Fofanov, Y., and SantaLucia, J. (1, 2011) Development and Characterization of a Highly Specific and Sensitive SYBR Green Reverse Transcriptase PCR Assay for Detection of the 2009 Pandemic H1N1 Influenza Virus on the Basis of Sequence Signatures. *Journal of Clinical Microbiology,* **49**, 335–344.

[36] Pierce, K. E., Khan, H., Mistry, R., Goldenberg, S. D., French, G. L., and Wangh, L. J. (11, 2012) Rapid detection of sequence variation in Clostridium difficile genes using LATE-PCR with multiple mismatch-tolerant hybridization probes. *Journal of Microbiological Methods,* **91**, 269–275.

[37] Moser, M. J., Christensen, D. R., Norwood, D., and Prudent, J. R. (2, 2006) Multiplexed Detection of Anthrax-Related Toxin Genes. *The Journal of Molecular Diagnostics,* **8**, 89–96.

[38] Ronish, B., Hakhverdyan, M., Ståhl, K., Belák, S., LeBlanc, N., and Wangh, L. (3, 2011) Design and verification of a highly reliable Linear-After-The-Exponential PCR (LATE-PCR) assay for the detection of African swine fever virus. *Journal of Virological Methods,* **172**, 8–15.

[39] Zhang, D. Y., Chen, S. X., and Yin, P. (2012) Optimizing the specificity of nucleic acid hybridization. *Nature Chemistry,* **4**, 208–214.

[40] Tan, Z.-J. and Chen, S.-J. (2, 2006) Nucleic Acid Helix Stability: Effects of Salt Concentration, Cation Valence and Size, and Chain Length. *Biophysical Journal,* **90**, 1175–1190.

[41] Cha, R. S. and Thilly, W. G. (12, 1993) Specificity, efficiency, and fidelity of PCR.. *Genome Research,* **3**, S18–S29.