applied sciences

MDPI

Article

A Method for Degenerate Primer Design Based on Artificial Bee Colony Algorithm

Ruhui Liu, Jiaxu Ning *, Yueqiu Jiang *, Xianghe Wang and Jiaxuan Wu

School of Information Science and Engineering, Shenyang Ligong University, Shenyang 110159, China; liuruhui_19990716@163.com (R.L.); wangxianghe1998@126.com (X.W.); wujiaxuan_will@163.com (J.W.)
* Correspondence: ningjiaxu@sylu.edu.cn (J.N.); yueqiujiang@sylu.edu.cn (Y.J.)

Abstract: Aiming to address the complex degenerate primer design problem in the biological field, in this paper, we design a degenerate primer optimization model considering primer coverage and degeneracy that allows a small number of base mismatches, and propose a global optimization method based on the artificial bee colony algorithm. The proposed algorithm combines the idea of the ant colony algorithm with the optimization process of the artificial bee colony algorithm, overcomes the disadvantage of the uncertain candidate solution length of the artificial bee colony algorithm in solving discrete optimization problems, designs the search space model according to the construction process of candidate solution in ant colony optimization algorithm, and redesigns various bee foraging strategies according to the optimization process information. In the comparative experiments on DNA template sequences of different scales, the degenerate primer designed by the proposed algorithm is superior to the existing methods in terms of stability, specificity, coverage and degeneracy.

Keywords: degenerate primer design; artificial bee colony algorithm; discrete optimization; search space model; foraging strategies

Citation: Liu, R.; Ning, J.; Jiang, Y.; Wang, X.; Wu, J. A Method for Degenerate Primer Design Based on Artificial Bee Colony Algorithm. *Appl. Sci.* **2022**, *12*, 4992. https://doi.org/10.3390/app12104992

Received: 6 April 2022
Accepted: 13 May 2022
Published: 15 May 2022

1. Introduction

In recent years, degenerate polymerase chain reaction (PCR) technology has been widely used in new gene cloning, gene expression detection, virus detection and genome research. It has the advantages of rapidity, simplicity, and high sensitivity [1]. A PCR primer sequence is called degenerate if some of its positions have several possible bases. The degeneracy of the primer is the number of unique sequence combinations it contains. This can improve the coverage of DNA template, but values that are too high will reduce the specificity of PCR. Therefore, the suitability of degenerate primers directly affects the success rate of PCR.

The design of degenerate primers can be described as a given set of n strings and integers k, d and m, looking for a primer of length k and degeneracy of at most d that matches at least m input strings [2,3].

In traditional primer design, if primers are required to match the target DNA sequence and some unknown sequences as much as possible, the number of bases at each position along the sequence is typically counted, and the appropriate primer length to minimize its length determined [4]. However, this method is often used, due to the inconsistency between the sequence and inappropriate fitness evaluation function, during operations in primers with higher degeneracy. Therefore, in the process of primer design, primers should usually be selected to match some but not all of the target DNA sequence.

In recent years, iterative heuristic algorithms have been widely used to solve primer design problems. Treeratanajaru W et al. [5] proposed a method using dynamic pattern matching to complete the design of degenerate primers, inputting nucleotide sequences from different bacteria into a system consisting of three steps: data reconstruction, primer

design and attribute filtering, freely combining the results with Gibbs, and designing and selecting the most suitable sequence as a series of primers. Balla S et al. [6] proposed an algorithm for designing the minimum degeneracy of degenerate primers for a given DNA sequence that combined the topic discovery methodology and iterative technology to perform testing and comparison in random and real biological data sets. Souvenir R et al. [7] proposed an iterative beam search algorithm and multiple iterative primer selector for the design of multiple degenerate primers that was superior to existing related algorithms for degenerate primer design and had a smaller number of stray amplifiers. Wu J S et al. [8] proposed the use of the genetic algorithm to solve the primer design problem, and the constraint conditions of primer design were described using symbols and formulations; the algorithm was able to design primers meeting the requirements of specific restriction sites and specificity. Liang H L et al. [9] proposed a new method for designing multiple PCR primers using the genetic algorithm and the MAP model; this method found several sets of appropriate primer pairs for five gene family cDNA templates, which were not only able to meet many primer design constraints, but also made the primers specific. In addition, software systems based on iterative heuristic algorithms have been developed for the design of degenerate primers, such as Primer 3 [10], Primo Degenerate 3.4 [11], CODEHOP [12], etc. Linhart C et al. [13] introduced the design of maximum coverage degenerate primers and minimum degeneracy primers, and developed a program named HYDEN on the basis of the proposed approximate algorithm, successfully applying it for the identification of olfactory receptor genes in mammals. Cickovski T et al. [14] used existing algorithms to design cluster-based degenerate primers, and applied them in parallel in the GPUDePiCt software package using the shared memory model and graphics processing units (GPUs) to accelerate processing, and tested them in a large number of sequences in the human genome.

As can be seen from the above studies on degenerate primers, there are three cases addressed by research on degenerate primer design: maximum coverage degenerate primer design, that is, given a set of n strings, find a primer with length k to match as many strings as possible; minimum degeneracy degenerate primer design, that is, given a positive integer d, m and a set with n strings, find a primer with length k, make it match at least m strings and have the minimum degeneracy d; and design of minimum degeneracy primers that allow mismatches, that is, given positive integers d, e and a set with n strings, find a primer of length k that matches at least m strings, has minimum degeneracy d, and allows at most e mismatches. A finite number of mispairing experiments have little effect on the results and can also improve the amplification efficiency of primers.

Most of the existing methods are suitable for one or two cases of degenerate primer design, but it is difficult to solve more complex problems.

In addition, the swarm intelligence optimization algorithms [15,16] emerging in recent years have been demonstrated to be an effective means of solving complex optimization problems. Among them, the artificial bee colony algorithm offers good performance. Its application for solving practical problems in various fields has been attracting attention [17–20]. Therefore, this paper studies a solution method for degenerate primer design based on the artificial bee colony algorithm.

The main contributions of this paper are as follows: (1) A method based on the artificial bee colony algorithm for solving degenerate primer design problems is proposed that can not only meet the constraints of primer design, but also allows mismatch while considering the coverage and degeneracy of primers. (2) Combining the optimization process of the artificial bee colony algorithm and the ant colony algorithm, an optimization model of the hybrid artificial bee colony algorithm is constructed. (3) Based on the idea of the ant colony algorithm, the foraging strategies of all kinds of bees are redesigned, and a search space model and pheromone matrix are constructed.

2. Problem Description

The aim of this paper is to find the degenerate primer with the minimum degeneracy and the maximum coverage among all possible candidate primers in the case of a small number of base mismatches. To facilitate evaluation, for any candidate primer P, the coverage and degeneracy were evaluated in this paper using Formula (1).

$$f(P) = \mu Unconverge\ (P) + \varphi Degeneracy\ (P) \tag{1}$$

In Formula (1), *Unconverge*() represents the uncovering function, that is, the number of primers that failed to match DNA template sequences. For example, if there are 10 groups of target DNA sequences and a candidate primer P completes matching with 6 groups of target sequences, the uncoverage is 4. *Degeneracy*() stands for a degeneracy function whose degeneracy is at most d. Degeneracy refers to a degenerate PCR primer as a primer sequence that contains several possible bases in one or more positions. For example, both DNA and primer sequences are composed of four bases of AGCT. Assuming that there is a primer P, P = {A}{AC}{GT}{ATG}, then the length k of P is 4 and the degeneracy d = $1 \times 2 \times 2 \times 3$ is 12. μ and φ are the corresponding weight parameters.

In addition, for a pair of candidate primers, the following constraints need to be met in this study:

In general, the primer length should be between 18 and 26 bps. The *length*() function is defined to represent the length of the primer, and its length is the sum of the number of bases at each position, as shown in Formula (2).

$$length(P) = \begin{cases} true & if\ 18 \le |P| \le 26 \\ false & Otherwise \end{cases} \tag{2}$$

where $|P|$ indicates the total number of bases in primer P.

The differential length of a primer pair is restricted to being smaller than 3 bps. $P_{forward}$ denotes the forward primers in a primer pair, and $P_{reverse}$ denotes the reverse primers in a primer pair. Define *lengd*() function to represent the length difference of primer pairs, as shown in Formula (3).

$$lengd(P) = \begin{cases} true & if\ \left||P_{forward}\right| - \left|P_{reverse}|\right| \le 3 \\ false & Otherwise \end{cases} \tag{3}$$

In general, the primer design has strict requirements on temperature. An empirical formula was proposed by Wallace for calculating the melting temperature of a primer P with a length between 18 and 26 bps. This function *Tm*() can be written as in Formula (4).

$$Tm(P) = (\#G + \#C) \times 4 + (\#A + \#T) \times 2 \tag{4}$$

This simple formula depends directly on the length and composition of the primer. #*G* indicates the amount of nucleotide "*A*" in *P*, #*T* indicates the amount of nucleotide "*T*" in *P*; #*C* and #*G* can then be defined accordingly. Additionally, the differential melting temperature of a primer pair must be under 2 °C. The *Tmd*() function is as shown in Formula (5).

$$Tmd(P) = \begin{cases} true & if\ \left|Tm(P_{forward}) - Tm(P_{reverse})\right| \le 2 \\ false & Otherwise \end{cases} \tag{5}$$

The *GC* content is the ratio of the number of nucleotide "*G*"s and the number of nucleotide "*C*"s in the primer *P* sequence. It should be limited to within a certain range. In

general, an appropriate range of *GC* content for a primer is between 40% and 60%. The *GC* content $GC(P)$ is as shown in Formula (6).

$$GC(P) = \frac{\#G + \#C}{|P|} \times 100\% \tag{6}$$

where #*G* indicates the number of nucleotide "*G*"s in *P*, and #C indicates the number of nucleotide "*C*"s in *P*. The $GC_{content}(P)$ function is as shown in Formula (7).

$$GC_{content}(P) = \begin{cases} true & if\ 40\% \leq GC(P) \leq 60\% \\ false & Otherwise \end{cases} \tag{7}$$

The 3′ end of a primer cannot choose nucleotide "*A*", it is better to choose nucleotide "*T*". When the 3′ end of a primer is mismatched, the last position is nucleotide "*A*", the chain synthesis can be triggered even under mismatched conditions. However, when the last position is nucleotide "*T*", mismatch initiation efficiency will be greatly reduced. The initiation efficiency of the nucleotide "*G*" and nucleotide "*C*" mismatch is between nucleotide "*A*" and nucleotide "*T*", so nucleotide "*T*" is the best choice for the 3′ end. See Figure 1.

Figure 1. The value of the 3′ end of the primer.

The *isGorC*() function is used to represent the 3′ end of the primer *P* with the nucleotides "*G*", "*C*", "*GC*" and "*CG*", which is as shown in Formula (8).

$$isGorC(P) = \begin{cases} true & if\ the\ end\ of\ 3' \ is\ G, C, GC, CG \\ false & Otherwise \end{cases} \tag{8}$$

Primers themselves should not have complementary sequences (no consecutive 4 bp complementarities), otherwise the primers themselves will fold into hairpin structures (as shown in Figure 2), thus affecting the combination of the primer and the template.

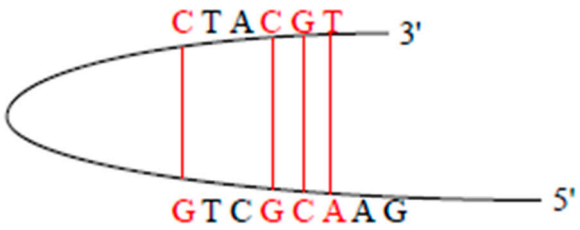

Figure 2. The card structure.

The *Sc*() function is used to indicate whether the primers themselves form a hairpin structure, as shown in Formula (9).

$$Sc(P) = \begin{cases} true & if\ there\ is\ no\ self-complementary\ between\ P_{forward}\ or\ P_{reverse} \\ false & Otherwise \end{cases} \tag{9}$$

There should also be no complementary sequences between the primer chains. In other words, there should be no complementarity between the forward primer and the reverse primer, and there should not be four consecutive bases of complementarity between the primer chains; in particular, the complementary overlap of the 3′ end should be avoided in order to prevent the formation of a primer double chain (or dimer). The *Pc*() function

was used to indicate whether double chain structures were formed between primer chains, as shown in Formula (10).

$$Pc(P) = \begin{cases} true & if\ there\ is\ no\ pair - completary\ between\ P_{forward}\ \text{and}\ P_{reverse} \\ false & Otherwise \end{cases} \quad (10)$$

3. The Proposed Algorithm

The aim of the proposed algorithm is to design degenerate primers. The degenerate primer design–hybrid artificial bee colony algorithm (DPD-HABC) is designed according to the construction process of candidate solutions and the use of pheromones.

The proposed algorithm mainly includes the representation of search space, updating of the pheromone matrix and the design of the foraging strategy. The model used to represent the search space defines the search space of the bee colony for the degenerate primer design problem and provides a representation of the food source. Since all kinds of bees share and exchange information through various pheromones in the process of foraging, the pheromone matrix and its update mode should be described first in the foraging process, before describing the foraging strategy. The representation of the search space, the updating of the pheromone matrix, and the design of the foraging strategy in this algorithm will be introduced in detail below.

3.1. Representation of the Search Space

For the degenerate primer design problem, each candidate solution is a primer pair, including a forward candidate primer and a reverse candidate primer, and the lengths of the forward primer and the reverse primer are between 18 and 26. Therefore, the search space for degenerate primer design can be expressed as a fully connected graph structure with 55 layers, as shown in Figure 3.

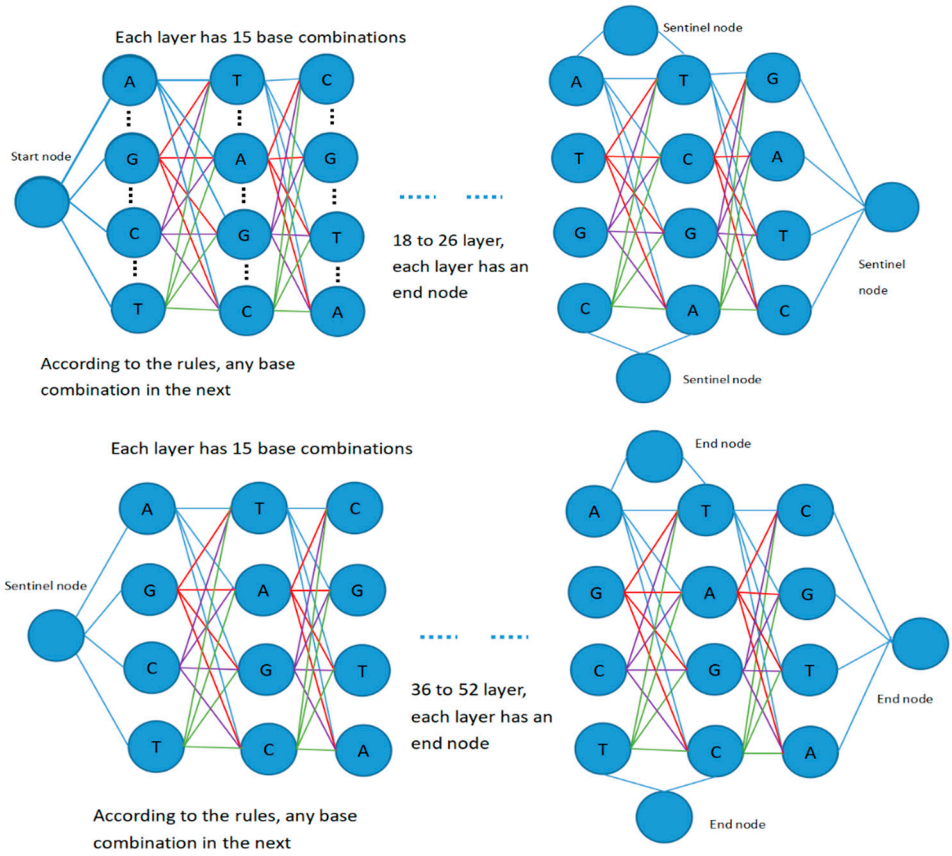

Figure 3. Representation of the search space.

The diagram contains four types of node: start node, intermediate node, sentinel node, and end node. The middle nodes of each layer are successively set as an effective combination of four base pairs, such as nucleotide "A", "T", "C", "G", "AT", "AC", "AG", "TC", "TG", "CG", "ATC", "ATG", "TCG", "ACG", "ATCG", etc. Sentinel nodes are mainly used to separate the forward primer and reverse primer, representing the end of the forward primer and the beginning of the reverse primer.

Setting up multiple sentinel nodes and end nodes is mainly performed to deal with the problem caused by variable primer length and to facilitate the completion of forward and reverse primer design. In each layer, appropriate base combination branches are selected according to the rules of base pairing and the constraints of coverage and degeneracy, and the cycle is iterated until appropriate primer pairs have been searched for and output results have been obtained or the maximum number of iterations has been reached. Then, the output results are filtered on the basis of primer design constraints, and the primer pairs and parameters of the forward and reverse primers that meet the requirements of the constraints are finally output.

In the process of constructing the search space, each bee constructs a candidate food source from the starting node in Figure 3, and calculates the selection probability of the next optional node in accordance with the probability state transition formula adopted by the next optional node according to its category (i.e., employed bee, onlooker bee, or scout bee). The next node to visit is then selected by means of a roulette game. For each bee, including the intermediate nodes in the next layer, several bases or base combinations that meet the requirements are screened on the basis of the input template DNA sequence and the base matching principle, and then the selection probability of the corresponding nodes is calculated. In this way, nodes are selected layer by layer until an end node is encountered. Therefore, in addition to the above 15 base combinations, an end node is also set at nodes 46 to 54, so that the construction of the candidate food sources can be completed at any time. In this way, each bee completes the construction process of a candidate food that completes the design of a candidate primer pair.

3.2. Update of the Pheromone Matrix

In order to construct candidate food sources with variable lengths, the pheromone concept employed in ant colony optimization is used in the artificial bee colony algorithm. The DPD-HABC algorithm contains two kinds of pheromone matrix: the global pheromone matrix and the local pheromone matrix.

The global pheromone matrix mainly identifies the preferred search area of the current employed bee group and represents information on the current group's search experience. Each employed bee sets up a local information matrix to record the individual search experience information it has obtained so far. In this way, each bee is able to forage according to the current search experience information of the whole population or the search experience information of individual employed bees. In the process of constructing candidate food sources, the greater the number of pheromones on a given side of the search space, the greater the probability of that side being selected.

In the DPD-HABC algorithm, when all employed bees have generated candidate food sources, the value of each element in the global pheromone matrix, that is, pheromone concentration, will be reduced in the same proportion to simulate pheromone volatilization. Then, *CBEST*, the best quality food source among these newly generated candidate food sources, is used to enhance the pheromones on the edges. For the pheromone τ_{ijk} on the edge between the j-th vertex of layer i and the k-th vertex of layer $i + 1$, the updated pheromone value is calculated as shown in Formula (11).

$$\tau_{ijk} \leftarrow (1 - \rho)\tau_{ijk} + \Delta\tau_{ijk}^{cbest} \tag{11}$$

where $\rho \in (0,1)$ represents the volatilization rate of the pheromone. If an edge arc is included in the *Cbest* path, set the increment $\Delta\tau_{ijk}{}^{cbest}$ of pheromone on the edge to the reciprocal of

the fitness function value of *CBEST*. In the process of algorithm solution, Formula (1) is used as the fitness function of candidate primers, and its calculation is shown as Formula (12).

$$\Delta\tau_{ijk}^{cbest} = \begin{cases} \frac{1}{f(CBSEST)} & if < (i,j),(i+1,k) > \in CBEST \\ 0 & otherwise \end{cases} \quad (12)$$

The updating of the local pheromone matrix for each employed bee is mainly carried out on the basis of the quality of the candidate food source constructed by the attracted onlooker bees. In this paper, the local pheromone matrix is updated using the best candidate path information generated by onlooker bees attracted by each employed bee in each evolution process. The calculation method for the value of each element in the local pheromone matrix is the same as that in the global pheromone matrix. In this paper, the pheromone values in the pheromone matrix change between [$\tau_{\min}$, $\tau_{\max}$], and the calculation method of $\tau_{\min}$ and $\tau_{\max}$ is the same as that in the MMAS algorithm [21].

Furthermore, for the current best food source F recorded by an employed bee, each update means that a better food source F′ has been found than F. By comparing the corresponding paths of these two food sources, we can identify what changes in F can be transformed into food source F′, and these changes are the fundamental reason for the food source becoming better. To this end, the DPD-HABC algorithm configures an identity matrix LMark of a binary type for each pheromone matrix to record side arcs contained in F′ but not contained in F.

3.3. Design of Foraging Strategy

According to the search space representation in Figure 3, each candidate food source corresponds to a path from the start node to the exit node. Therefore, the degenerate primer design problem can be treated as a path optimization problem, and ant colony optimization is an effective means for dealing with such problems.

In order to represent all kinds of bees in Figure 3 for effective search, in this paper, the concept of pheromones is introduced into the artificial bee colony optimization model in order to solve the problem of the variable lengths of food sources, and by means of the pheromone record, all kinds of bees are able to access quality information in the process of searching for the food source, guiding their subsequent foraging process. A global pheromone matrix is added to the DPD-HABC optimization model algorithm to record the current information on the search state and guide the subsequent foraging process of the employed bees. In addition, each employed bee is equipped with a pheromone matrix to record the state information of foraging around the food source and to guide the foraging process of onlooker bees attracted by the food source. The foraging strategies and pheromone renewal strategies of the employed, onlooker and scout bees will be introduced below.

In the DPD-HABC algorithm, each employed bee constructs a candidate food source in accordance with the current global pheromone matrix, the pheromones in its own saved local pheromone matrix, and heuristic information. Heuristic information is embodied in the form of heuristic factors, which can guide the selection of the next node. When employed bee l generates a candidate food source at the j-th node of layer i, the selection probability formula used to select node k of the next layer is calculated as shown in Formula (13).

$$P_{ij}^{k} = \begin{cases} \frac{[\tau^{g}{}_{ijk}+\tau^{l}{}_{ijk}]^{\alpha} * [\eta_{ijk}]^{\beta}}{\sum_{h \in allowed_h} [\tau^{g}{}_{ijh}+\tau^{l}{}_{ijh}]^{\alpha} * [\eta_{ijh}]^{\beta}} & if\ k \in allowed_h \\ 0 & Otherwise \end{cases} \quad (13)$$

where $\tau^{g}{}_{ijk}$ and $\tau^{l}{}_{ijk}$ are the pheromone values from the j-th node in the i-th layer to the k-th node in the next layer, respectively, stored in the global pheromone matrix and the local pheromone matrix of l; α, β respectively represent the corresponding weights of the

pheromone factor and the heuristic factor; $allowed_h$ is the set of nodes that can be accessed by the next layer; η_{ijk} is a heuristic factor. The calculation method is shown in Formula (14).

$$\eta_{ijh} = \frac{Converge_{ijh}}{Degeneracy_{ijh}} \tag{14}$$

where $Degeneracy_{ijh}$ represents the degeneracy of the current partial path after the j-th node of layer i is selected in the node h of the next layer, and $Converge_{ijh}$ represents the coverage of the current partial path after the j-th node of layer i is selected in the node h of the next layer. Because we are aiming for degenerate primer design with maximum coverage, when the degeneracy is certain, the larger the coverage, the better, and the value of the heuristic factor will increase accordingly.

Because the onlooker bee is mainly responsible for local searching around the food source, in the DPD-HABC algorithm, after each onlooker bee selects a given employed bee, the construction of its candidate food source is mainly based on the information in the local pheromone matrix saved by the employed bee. For an onlooker bee that chooses employed bee l for the follow-up search, when it is at the j-th node of layer i in the process of constructing the candidate food source, the selection probability formula used to select the node k of the next layer is calculated as shown in Formula (15).

$$P_{ij}^{k} = \begin{cases} 1 & if\ Lmark[i,j,k] = 1 \\ \frac{[\tau^{l}{}_{ijk}]^{\alpha} * [\eta_{ijk}]^{\beta}}{\sum_{h \in allowed_h} [\tau^{l}{}_{ijh}]^{\alpha} * [\eta_{ijh}]^{\beta}} & if\ k \in allowed\ i \\ 0 & Otherwise \end{cases} \tag{15}$$

where the meaning of each symbol and the calculation method of the heuristic factor are the same as those presented in Formula (14).

In the artificial bee colony algorithm, scout bees are mainly responsible for exploring in the whole empty frame of the search space, finding new food sources and guiding the optimization process out of local optimization. while the global pheromone matrix will reflect the preferred search area and search intensity of the current employed bee, the local pheromone matrix is able to reflect the preferred search area and search intensity of the observation bee. When the employed bee abandons the food source and turns to the onlooker bee for the search, this means that the employed bee thinks it is difficult to find a better food source in these areas, and exploration should be directed towards other areas. Therefore, in the DPD-HABC algorithm, when the employed bee abandons the food source, a new candidate food source is generated in a region far from the employed bee's current preference by combining the global pheromone matrix with the local pheromone matrix. For the employed bee l, when it becomes a scout bee, when it is at the j-th node of the i-th layer in the process of generating a candidate food source, the selection probability formula used to select the node k of the next layer is calculated as shown in Formula (16).

$$P_{ij}^{k} = \begin{cases} \frac{[\tau_{\max} - (\tau^{g}{}_{ijk} + \tau^{l}{}_{ijk})/2]^{\alpha} * [\eta_{ijk}]^{\beta}}{\sum_{h \in allowed_h} [\tau_{\max} - (\tau^{g}{}_{ijh} + \tau^{l}{}_{ijh})/2]^{\alpha} * [\eta_{ijh}]^{\beta}} & if\ k \in allowed_h \\ 0 & Otherwise \end{cases} \tag{16}$$

where τ_{max} represents the upper bound of the value of the pheromone in the pheromone matrix. When a new candidate food source is constructed by roulette according to the probability formula, the scout bee turns into an employed bee in order to search again, takes the newly generated candidate food source as the current best food source, and reinitializes the local information matrix by flipping it, that is, making any element in the local pheromone matrix $\tau^{l}{}_{ijk} = \tau_{\max} - \tau^{l}{}_{ijk}$.

Evaluating food sources is an essential step for the DPD-HABC algorithm when solving the problem of degenerate primer design. Here, the food sources are primer pairs.

When comparing the food sources, this paper divides them into three cases: the two compared food sources meet the constraints; one of them satisfies the constraints, while the other does not satisfy the constraints; and neither meets the constraints.

3.4. Algorithm Description

In order to solve the problem of degenerate primer design, the DPD-HABC algorithm includes four processes: initialization, employed bee foraging, onlooker bee foraging, and scout bee foraging. In the initialization stage, the pheromone matrix and the identity matrix contained in the algorithm are initialized. The value of each element in the pheromone matrix is set to the maximum value, max, and each element in the identity matrix is set to 0. of the method for calculating the probability of attracting onlooker bees and the method for allocating onlooker bees are the same as in the standard artificial colony algorithm. The specific process of the algorithm is shown in Algorithm 1.

Algorithm 1. DPD-HABC Algorithm for Degenerate Primer Design

Input
f: the fitness evaluation function
Ω:constraint condition of degenerate primer design
Output
S_{bs}: the optimal primer pairs
1. **Begin**
//Initialization process
2. Initialize food sources and pheromone trails and heuristic information
3. **repeat**
//employed bee foraging process
4. **for** each employed bee **do**
5. Initialize its local phoneme matrix and the Lmark matrix
6. Generate a new candidate food source S and evaluate it
7. Construct a complete primer pairs S generated by Equation (13)
8. **if** f(S) < f(S_{bs}) **then**
9. $S_{bs} \leftarrow S$
10. **end if**
11. **end for**
//onlooker bee foraging process
12. **for** each onlooker bee do
13. produce the new solutions v_i from the solutions S_{bs} generated by Equation (15)
14. apply the greedy selection process between v_i and S_{bs}
15. end for
//scout bee foraging process
16. for each scout bee do
17. if existing the abandoned solution S then
18. Send scouts based on the scout foraging strategy generated by Equation (16)
19. end if
20. end for
21. for each component i in graph do
22. $\tau_{ij} \leftarrow$ update pheromone
23. end for
24. Memorize the best food source positions (solutions)
25. until maximum fitness evolution number reached or other termination condition met
26. return S_{bs}
27. End

The analysis of the time complexity and space complexity of the DPD-HABC algorithm for solving the problem of degenerate primer design is as follows: O($n*m$) operations are required when the employed bees conduct a neighborhood search, based on the location of the food source in their memory, to find a better food source attached to the food source; the onlooker bees choose a certain food according to the dance of the employed bees in

the hive, and it takes O($n*m$) operations to calculate the probability formula. It takes O(n) operations for the scout bee to randomly select a food item. Thus, the total time complexity can be calculated as shown in Formula (17).

$$T(n) = O(N_c * n^2 * m) \tag{17}$$

where N_c is the number of cycles, n is the number of base pair combinations run, and m is the total number of bees. Thus, the time complexity increases with increasing DNA template size. The total space complexity is shown in Formula (18).

$$S(n) = O(n^2) + O(n * m) \tag{18}$$

4. Experimental Results

In this paper, 100 sets of DNA template sequences were used, most of which came from real molecular biology experiments. All experiments using this algorithm were completed using the Eclipse platform on the same PC (3.40 GHz CPU and 16 GB RAM) and implemented in the Java development programming language.

4.1. Parameter Adjustment

Parameter α represents the corresponding weight of the pheromone factor in the probability state transition formula, which is usually set to 2. Parameter β represents the weight of the corresponding heuristic factor in the formula, and its value is related to the degeneracy and complexity of the primer pairs, and is usually set to 8. Parameter r is the pheromone volatilization coefficient, which is usually set to 0.02. Parameters μ and φ are the key points discussed in this section, representing the corresponding weights of coverage and degeneracy, respectively. By analyzing the number of constraint violations in the process of degenerate primer design, it was found that the value range of parameter μ was between [5, 25], because the number of constraint violations was the lowest within this range; on the other hand, the value range of parameter φ was roughly within [1, 5]. Similarly, because the number of constraint violations was the lowest within this range, the designed primer pair was the best. The values of the above parameters are shown in Table 1.

Table 1. Parameter settings.

Parameter Type	Value Range
α	2
β	8
ρ	0.02
μ	[5, 25]
φ	[1, 5]

To further determine the specific values of parameters m and j within this range, we performed related experiments, and the final experimental results are presented in the form of box diagrams, displayed in Figures 4 and 5, where the abscissa represent the value ranges of the parameters, and the ordinates represent the number of constraint violations after standardization using a given value.

From Figure 4, on the basis of the analysis and comparison of the box line diagrams for four typical data sets, it could be found that when the value of the parameter m was 10, the distribution of the box line diagram was more uniform. In comparison, the minimum value at this time was the best. For example, in case 2, where m = 10, it was obvious that the maximum, upper quartile, median, lower quartile and minimum values were better than those of the other cases with other parameter settings. Therefore, the value of the parameter m should be set to 10.

In Figure 5, on the basis of an analysis and comparison of the box graphs of four typical data sets, it can be seen that when the value of the parameter φ was 3, the distribution of the

box graph was more uniform. In comparison, the lowest number of constraint violations is the best. Therefore, the value of the parameter φ should be set as 3.

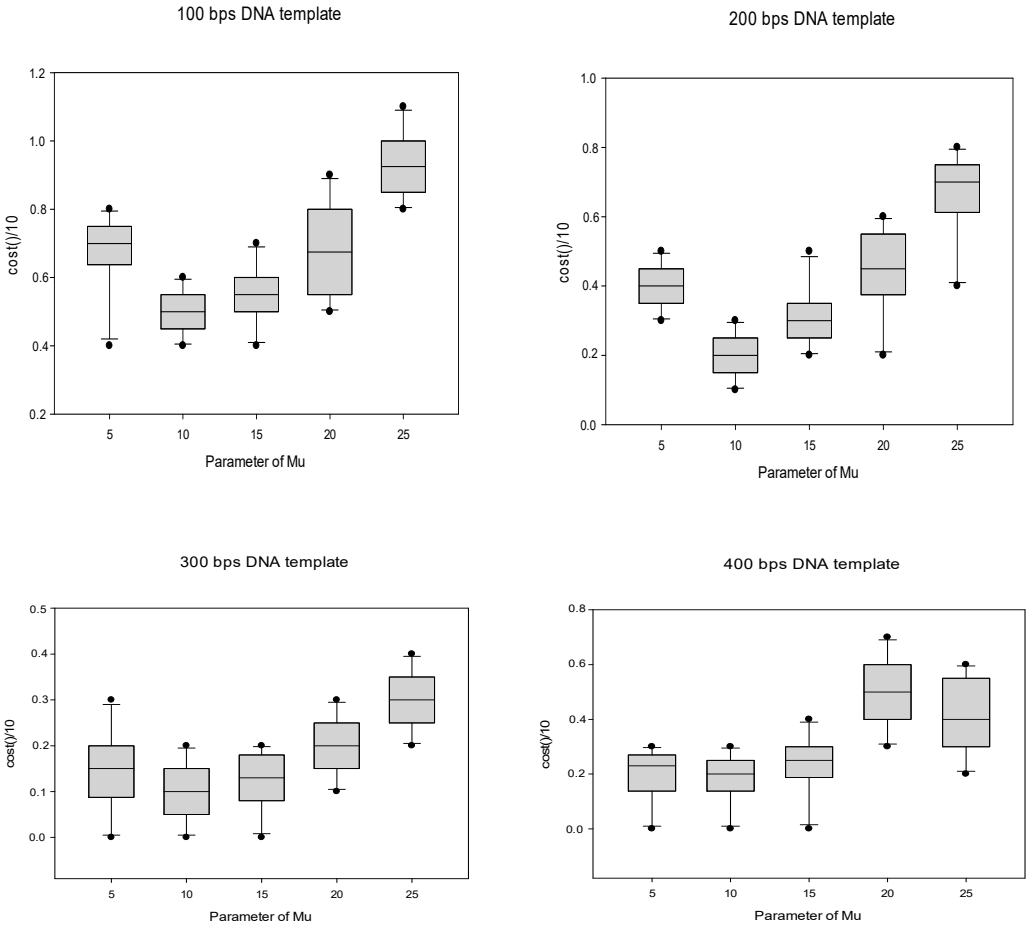

Figure 4. The adjustment of the parameter μ.

4.2. Comparison with Related Algorithms

In order to make the experimental data more fair and reasonable, the experimental section compared the experimental results of DPD-HABC algorithm for the design of degenerate primers with the results of the relatively mature software programs Primer 5 [22] and Oligo 7 [23]. The experimental results of this paper will mainly be described from two different perspectives: convergence analysis and comparative and descriptive statistical analysis.

4.2.1. Analysis and Comparison on the Basis of Convergence

In the convergence curve, the abscissa axis represents the fitness evaluation time, and the ordinate axis represents the value of the fitness evaluation function after standardization over a certain fitness evaluation time. The convergence curve of the DPD-HABC algorithm and the software Primer 5 and Oligo 7 for solving degenerate primer design on 500 bps and 600 bps DNA templates, respectively, is shown in Figure 6.

From Figure 6, it can be seen that on the 500 bps DNA template sequence, Primer 5 converges fastest when analyzed from the perspective of convergence; however, for the same fitness evaluation time, the value of the fitness evaluation function of the DPD-HABC algorithm is smaller, so the performance of the DPD-HABC algorithm is better.

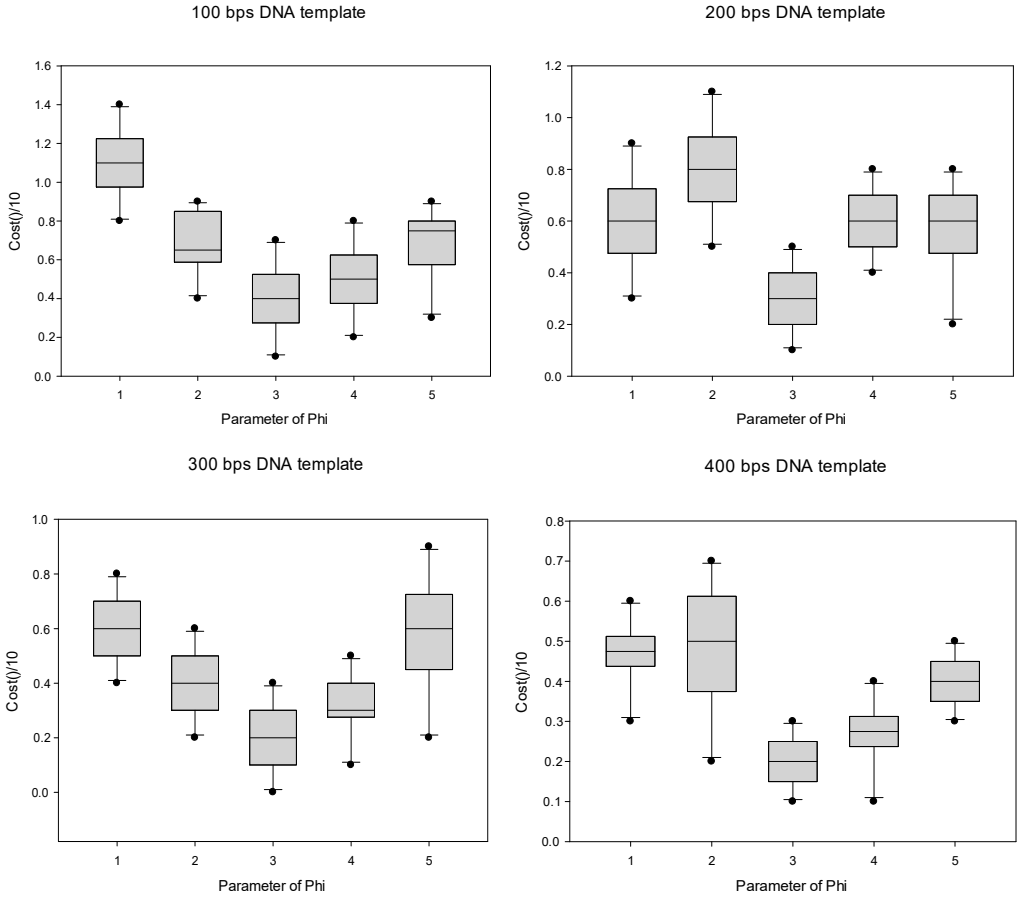

Figure 5. The adjustment of parameter φ.

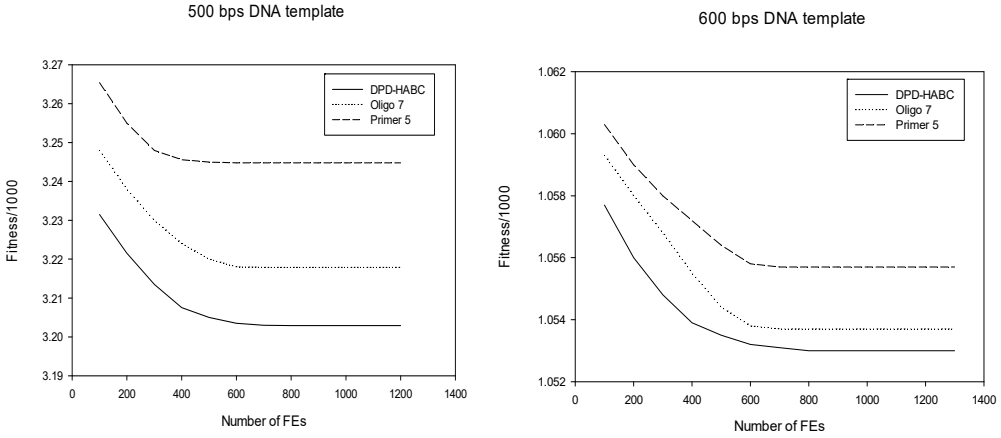

Figure 6. Convergence graphs for DPD.

4.2.2. Analysis and Comparison on the Basis of Descriptive Statistics

From 100 groups of DNA template combinations, the 500 bps DNA template and the 600 bps DNA template are selected as the test cases. The specific experimental data (such as primer pair composition, degeneracy, coverage, primer length, GC content, specificity, and other evaluation indexes) of the DPD-HABC algorithm, Primer 5, and Oligo 7 are shown in Table 2.

Table 2. Comparison of three methods on 500 bps DNA template.

	DPD-HABC	Primer 5	Oligo 7
Forward primer (5′-3′)	T{AC}{AT}{AGC}A {AG}A{GT}GG G{TC}{TG}{TC}{GT} {AC}AGG{TC}	{AT} A {AT}{AGC}{CA} GC{AT}{GT}{GC} {TG}T{TC}{TG}C T{AC}{AT}GG	{AT}{AC}{AT}A{CA} {AG}{GC}A{GT}G {CTG}G{TC}{TG}{TC} T{AC}AG{AG}
Reverse primer (5′-3′)	TAC{GC}{ATC} {AT}{GC}{TC}C A { AG } {AT}GC{TG} {AT}{AT}TC{CT}	{AT}{AC}C{GC}{ATC} {GA}{AT}G{TC}C {TG}G{AT}{AG}C {CG}{AT}{AT}TC	{AG} A CG{AT} {GAT}AG{TC}{CG} TG{AT}A{AC}
Degeneracy (F/R)	3072/3072	6144/12,288	6144/4608
Converge (F/R)	30	20	25
Primer length (F/R)	20/20	20/20	20/20
Melting temperature (F/R) (°C)	57/55	60/62	61/63
Temperature difference (°C)	2	2	2
GC-content (F/R) (%)	52	53	54
GC-clamp (F)	CG	CG	GG
GC-clamp (R)	G	G	CG
Self-annealing	No	No	No
Pair-annealing	No	No	No
specificity	Yes	No	No

From Tables 2 and 3, it can be seen that, from the perspective of degeneracy, the degeneracy of the DPD-HABC algorithm based on the hybrid artificial bee colony optimization model for degenerate primer design is 3072 and 1536, respectively; values which are significantly lower than the results for the solutions obtained when using the software programs Primer 5 and Oligo 7. With respect to coverage, the coverage of the DPD-HABC algorithm based on a hybrid artificial bee colony optimization model for degenerate primer design is 30, which is significantly higher than that of Primer 5 and Oligo 7. In addition, the primer pairs solved using the DPD-HABC algorithm are specific compared with the other two software programs, and the quality of the primer pairs obtained using the DPD-HABC algorithm are inseparable from the processing of search space. The specific experimental data of the DPD-HABC algorithm on two DNA template sequences of 100 bps and 200 bps are shown in Table 4.

Table 3. Comparison of three methods on a 600 bps DNA template.

	DPD-HABC	Primer 5	Oligo 7
Forward primer (5′-3′)	T{AG}G{CT}{AC} {AT}G{AT}{AC} C{CT}{GT}G{TAC} GC{GA}CG	{AT}{AG}{AG}{CT}{GAC}{AT}G{AT}A C{CT}{GT}G{TAC} {GA}C{GA}CG	{AT}G{AG}C{AC} AG{ATG}{AG} CT{GT}{AG}{TC} G{TC}{GAC}CG
Reverse primer (5′-3′)	CAG{AC}{GA} C{AT}{GT}{AC} {AC}TG{AG}{TAC} A{CT}AG{AG}	CA{AG}{AC}{GA} C{AT}{GT}{AC} {AC}{AGT}G{AG}{TAC} A{CT}A{GAT}G	CA{AG}CA C{AT}T{AGC} {AC}{AT}G{AG}{TC} {AT}{CT}A{GA}{GT}
Degeneracy (F/R)	1536/1536	9216/13,824	2304/3072
Converge (F/R)	30	20	26
Primer length (F/R)	19/19	19/19	19/19
Melting temperature (F/R) (°C)	54/56	58/62	58/59
Temperature difference (°C)	2	2	2
GC-content (F/R) (%)	50	48	52
GC-clamp (F)	G	G	CG
GC-clamp (R)	C	CG	GC
Self-annealing	No	No	No
Pair-annealing	No	No	No
specificity	Yes	No	No

Table 4. Experimental results on 100 bps and 200 bps DNA templates.

	100 bps	**200 bps**
Forward primer (5′-3′)	{GC}GT{AT}GTA{TC}{AC}T CC{TG}{TG}AA{TG}{TG}GG	GA{TG}{TG}{AG}AGTG{AG} {GC}A{GA}GT{TG}{AG}TG
Reverse primer (5′-3′)	TG{AG}C{GT}T{CG}C A{CG} {AC}T{AG}G{AG}{AT}T{GT}GC	A{AC}TTAC{GT}{TG}{AT}G A{AC}GG{CT}{TG}A{TC}G
Degeneracy (F/R)	256/512	256/256
Converge (F/R)	30	30
Primer length (F/R)	20/20	19/19
Melting temperature (F/R) (°C)	54/55	56/54
Temperature difference (°C)	1	2
GC-content (F/R) (%)	53	51
GC-clamp (F)	GC	G
GC-clamp (R)	GG	G
Self-annealing	No	No
Pair-annealing	No	No
specificity	Yes	Yes

The specific experimental data of the DPD-HABC algorithm on two DNA template sequences consisting of 300 bps and 400 bps are shown in Table 5. On the basis of Table 5, it can be seen that when solving template sequences of different sizes, the degeneracy of the DPD-HABC algorithm is also different. With increasing template size, there is a corresponding increase in degeneracy. For example, when solving the 200 bps DNA template sequence, the degeneracy is 256; however, when solving the 300 bps DNA template sequence, the degeneracy is 1536. However, the coverage is not related to the size of the DNA template sequence. For example, when solving four groups of typical DNA template sequences, the coverage is 30. This is related not only to the number of bases in the template, but also to the excellent characteristics of the global search when using the DPD-HABC algorithm.

Table 5. Experimental results for the 300 bps and 400 bps DNA templates.

	300 bps	**400 bps**
Forward primer (5′-3′)	A{AC}T{GT}C{CA}T{AC} A{AG} G{CA}{AG}{GTC}T{GA}G{AG}G	T{ATC}AA{TC}G{AG}{CA}T{CA} {CA}T{GCA}AGC{AG}{GA}G
Reverse primer (5′-3′)	{ATG}{AT}TG{TC}C{TG}A{AC}A {AG}A{CG}{TC}GA{GA}G{AG}G	{TGC}C{AG}{AG}{TC}{GT}{TC}A GCA{CT}A{CA}CAC}A{AG}GG
Degeneracy (F/R)	1536/1536	1152/1536
Converge (F/R)	30	30
Primer length (F/R)	19/20	19/20
Melting temperature (F/R) (°C)	57/56	55/54
Temperature difference (°C)	0	1
GC-content (F/R) (%)	52	51
GC-clamp (F)	GG	G
GC-clamp (R)	C	GG
Self-annealing	No	No
Pair-annealing	No	No
specificity	Yes	Yes

5. Conclusions

Degenerate primer design is a complex problem. The candidate primers solved in this paper should not only be able to meet the requirements of multiple constraints, but also make the coverage of primers as large as possible and the degeneracy as low as possible under the condition of allowing a small number of base mismatches. Therefore, this paper establishes a degeneracy primer design optimization model, and puts forward a method for solving this problem that is based on the idea of artificial bee colony optimization. The

experimental results show that the primer pairs designed using this method not only meet the constraints and specificity requirements, but also improve the solution efficiency to a certain extent. In addition, the algorithm proposed in this paper is sensitive to parameter design, and it is difficult to find optimal parameter settings for specific test problems. In the evaluation of candidate primers, degeneracy and coverage are transformed into a single index that is scored by weighting, which requires the corresponding weight value to be provided prior to calculation, while the determination of the optimal weight value needs more pre-experiments in practical use. Therefore, the manner in which a multi-objective optimization model for degeneracy primer design can be established will be considered in future research, along with the design of an effective solution method that is able to furnish multiple candidate degeneracy primer design schemes that are simultaneously not dominated by each other.

Author Contributions: Conceptualization, Y.J. and J.N.; methodology, R.L. and J.N.; software, X.W.; validation, Y.J.; formal analysis, J.W.; investigation, R.L.; resources, J.N.; data curation, Y.J. and X.W.; writing—original draft preparation, R.L. and J.N.; writing—review and editing, R.L. and J.N.; visualization, J.W.; supervision, J.N.; project administration, Y.J. and J.N.; funding acquisition, Y.J. and J.N. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Liaoning Province Doctoral Research Start-up Fund project, grant number 2020-BS-152. This work was supported by the Xingliao talent plan, grant number XLYC1902095. This work was supported by Scientific Research Young Talents Project of Liaoning Education Department, grant number LJKZ0266.

Data Availability Statement: The data used to support the findings of this study are available from the corresponding author upon request.

Acknowledgments: Thanks to the reviewers and editors for their constructive suggestions, which have been very useful for improving this paper.

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

References

1. Shi, Z.X.; Wang, H.L.; Su, G.F.; Huang, L.Y. Theory and application of degenerate PCR. *Biotechnol. Commun.* **2004**, *2*, 172–175.
2. Linhart, C. The Degenerate Primer Design Problem. Master's Thesis, School of Computer Science, Tel Aviv University, Tel Aviv, Israel, November 2002. Available online: http://www.cs.tau.ac.il/~{}chaiml/biology/dpd_thesis.ps.gz (accessed on 6 May 2022).
3. Linhart, C.; Shamir, R. The degenerate primer design problem: Theory and applications. *J. Comput. Biol.* **2005**, *12*, 431–456. [CrossRef] [PubMed]
4. Singh, V.K.; Mangalam, A.K.; Dwived, S.; Naik, S. Primer premier: Program for design of degenerate primers form a protein sequence. *BioTechniques* **1998**, *24*, 318–319. [CrossRef] [PubMed]
5. Treeratanajaru, W.; Watcharamul, S.; Lipikorn, R. Degenerate primer design system for gene biodiversity study using dynamic pattern matching. In Proceedings of the 2012 7th International Symposium on Health Informatics and Bioinformatics (HIBIT), Nevsehir, Turkey, 19–22 April 2012; pp. 102–106.
6. Balla, S.; Rajasekaran, S. An efficient algorithm for minimum degeneracy primer selection. *IEEE Trans. Nanobiosci.* **2007**, *6*, 12–17. [CrossRef] [PubMed]
7. Souvenir, R.; Buhler, J.; Stormo, G. Selecting degenerate multiplex PCR primers. In *International Workshop on Algorithms in Bioinformatics*; Springer: Berlin/Heidelberg, Germany, 2003; pp. 512–526.
8. Wu, J.S.; Lee, C.; Wu, C.; Shiue, Y.L. Primer design using genetic algorithm. *Bioinformatics* **2004**, *20*, 1710–1717. [CrossRef] [PubMed]
9. Liang, H.L.; Lee, C.; Wu, J.S. Multiplex PCR primer design for gene family using genetic algorithm. In Proceedings of the 7th Annual Conference on Genetic and Evolutionary Computation, Washington, DC, USA, 25–29 June 2015; ACM: New York, NY, USA, 2005; pp. 67–74.
10. Primer3web. 2022. Available online: https://primer3.ut.ee (accessed on 6 May 2022).
11. Primo Degenerate 3.2. 2022. Available online: https://www.changbioscience.com/primo/dhowto.html (accessed on 6 May 2022).
12. CODEHOP. 2022. Available online: https://blocks.fhcrc.org/codehop.html (accessed on 6 May 2022).
13. Linhart, C.; Shamir, R. HYDEN—A Software for Designing Degenerate Primers. 2003. Available online: http://www.cs.tau.ac.il/~{}rshamir/hyden (accessed on 6 May 2022).
14. Cickovski, T.; Flor, T.; Irving-Sachs, G.; Novikov, P.; Parda, J.; Narasimhan, G. GPUDePiCt: A parallel implementation of a clustering algorithm for computing degenerate primers on graphics processing units. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2015**, *2*, 445–454. [CrossRef] [PubMed]

15. Zhao, H.; Zhang, C.; Zheng, X.; Zhang, C.; Zhang, B. A decomposition-based many-objective ant colony optimization algorithm with adaptive solution construction and selection approaches. *Swarm Evol. Comput.* **2022**, *68*, 100977. [CrossRef]
16. Liu, A.; D Jiang, D.; Zhang, C.; Zhao, H.; Zhao, Q.; Zhang, B. A Novel Fireworks Algorithm for the Protein-Ligand Docking on the AutoDock. *Mob. Netw. Appl.* **2021**, *2*, 657–668. [CrossRef]
17. Os, A.; Ba, A.; Dka, B. Archive-based multi-criteria Artificial Bee Colony algorithm for whole test suite generation. *Eng. Sci. Technol. Int. J.* **2021**, *24*, 806–817.
18. Kumar, M.S.; Rajamani, D.; Nasr, E.A.; Balasubramanian, E.; Mohamed, H.; Astarita, A. A Hybrid Approach of ANFIS—Artificial Bee Colony Algorithm for Intelligent Modeling and Optimization of Plasma Arc Cutting on Monel 400 Alloy. *Materials* **2021**, *14*, 6373. [CrossRef] [PubMed]
19. Zhao, H.; Zhang, C. A decomposition-based many-objective artificial bee colony algorithm with reinforcement learning. *Appl. Soft Comput.* **2020**, *86*, 105879. [CrossRef]
20. Wu, Y.; Xu, J.; Zhang, C. *A Heuristic Scout Search Mechanism for Artificial Bee Colony Algorithm*; Springer: Cham, Switzerland, 2019.
21. Stützle, T.; Hoos, H.H. MAX-MIN ant system. *Future Gener. Comput. Syst.* **2000**, *16*, 889–914. [CrossRef]
22. Primer Premier: Software for PCR Primer Design. 2022. Available online: http://www.premierbiosoft.com/primerdesign/ (accessed on 6 May 2022).
23. Rychlik, W. OLIGO 7 Primer Analysis Software. In *PCR Primer Design*; Methods in Molecular Biology™; Humana Press: Totowa, NJ, USA, 2007; p. 402.