Multiplex PCR Primer Design for Gene Family Using Genetic Algorithm

Hong-Long Liang
Department of Computer Science and Engineering
National Sun Yat-Sen University, Kaohsiung, Taiwan
70 Lien-hai Rd. Kaohsiung 804, Taiwan
886-7-525-4335
m9134619@student.nsysu.edu.tw

Chungnan Lee
Department of Computer Science and Engineering
National Sun Yat-Sen University, Kaohsiung, Taiwan
70 Lien-hai Rd. Kaohsiung 804, Taiwan
886-7-525-4335
cnlee@mail.cse.nsysu.edu.tw

Jain-Shing Wu
Department of Computer Science and Engineering
National Sun Yat-Sen University, Kaohsiung, Taiwan
70 Lien-hai Rd. Kaohsiung 804, Taiwan
886-7-525-4335
wujs@mail.cse.nsysu.edu.tw

ABSTRACT

The multiplex PCR experiment is to amplify multiple regions of a DNA sequence at the same time by using different primer pairs. Designing feasible primer pairs for multiplex PCR is a tedious task since there are too many constraints to be satisfied. In this paper, a new method for multiplex PCR primer design strategy using genetic algorithm is proposed. The proposed algorithm is able to find a set of suitable primer pairs more efficient and uses a MAP model to speed up the examination of the specificity constraint that is important for gene family sequences. The dry-dock experiment shows that the proposed algorithm finds several sets of primer pairs of gene family sequences for multiplex PCR that not only obey the design properties, but also have specificity.

Categories & Subject Descriptors

J.3 [**Computer Applications**]: Life and Medical Sciences-Biology and genetics

General Terms: Experimentation.

Keywords: Genetic Algorithm, Multiplex PCR, Primer Design

1. INTRODUCTION

In the advent of genetic engineering, the researches in the biological sciences on genetic analysis become popular and important. The main problem of the genetic analysis is how to duplicate specific and mass genes from a genome. The Polymerase Chain Reaction (PCR) has solved the troublesome requirement in medicine and in molecular biology [1], and also extended to many widespread applications. Before performing the PCR experiment, it is important to find a good primer pair. The choice of primers affects the success of PCR experiment. The PCR primer design needs to satisfy many constraints, such as the specificity, the primer length, the melting temperature and the GC content and so on. Designing a suitable PCR primer pairs is a hard work, since many constraints affecting the success or failure and of experiment most of these constraints are related to each other. To solve the problem, the primer design can be done by utilizing powerful computers and a suitable algorithm.

In order to analyze various kinds of DNA sequences at the same time, the multiplex PCR is proposed. It uses several suitable primer pairs to amplify several different lengths of the target sequences simultaneously in a test tube. Hence it is able to save the cost of time and money, produce quantification of DNA fragment and obtain more information in one PCR experiment. The multiplex PCR is used in many areas. For example, it is able to increase the speed of DNA sequencing, make decoding the genetic codes of genome more effective and faster. It is also very useful to detect the mutation of the genome that causes the disease

There are many researches or softwares for assisting primer design, but most of them only for single PCR primer design. Researches for multiplex PCR primer design are still only a few [2][3]. Nicodeme et al. proposed a two-step algorithm to design primer pairs for the multiplex PCR [4][5]. In the first step, they use the program PRIMER to select possible primer pairs to be candidates according to each target DNA sequence. And then, it uses an urn model to solve the problem of compatibility between primers, and tests melting temperatures and complementarity between any two primers of these candidates by using the program MULTIPCR. At last a set of suitable primer pairs from these candidates are selected for the multiplex PCR. Schoske et al. [6] presented a strategy includes designing, testing and optimization of multiplex PCR primer mixtures. They used the existing software Primer3 to select sets of primer pairs for single PCR and then select some candidates-primers by strictly examining the melting temperatures and complementarity constraints. At last, the comparisons between primer pairs are performed using their own algorithm. Based on the comparison, a set of suitable primer pairs are obtained. This approach gives a convenient way to obtain multiplex PCR primers and reduces the difficulty in multiplex PCR primer design by directly utilization publicly available software. However, it is a time consuming process and doesn't optimize the objective function simultaneously for multiplex PCR primer pairs. Furthermore, these methods take a point of assessing melting temperatures and complementarity, but some important constraints, like the specificity, may not be considered.

GECCO'05, June 25-29, 2005, Washington, DC, USA.

Kämpke et al. [7] proposed an algorithm using dynamic programming (DYNPA-APP2) to find multiplex PCR primer pairs. Although the approach can find a set of primer pairs simultaneously, some constraints like the specificity, the GC clamp, and so on are still not considered yet. In order to reduce the computational complexity, this method removes some potential solutions that may take a long time to compute. Hence, new algorithms and more complete constraints are needed to improve the multiplex PCR primer design.

Genetic algorithm (GA) is one of effective optimization search methods. The basic concept of GA was first proposed by John Holland in the 1970 [8]. Wu et al. [9] use GA for solving the single PCR primer design problem. They use the location of primer-pair in DNA sequence as individuals, and further performed the GA operators to the individuals. According to the results, this approach obtains a good solution. However, this method is still not able to find the multiplex PCR primer-pairs at one time.

Given a set of n DNA sequences and target length $|Target_i|$, where i=1,…,n, are amplified for each DNA sequence respectively. The goal of this paper is to find a set of n feasible primer pairs P_{pair_i} comprises n forward primers and n reverse primers, for every i=1,…,n, to satisfy all the constraints including self constraints for a single primer pair and inter-constraints for different pairs of primer.

The rest of this paper is organized as follows. In Section 2, the definitions of these constraints on multiplex PCR are presented. Main framework of the proposed GA is presented in Section 3. Dry dock experiments are conducted in Section 4. The conclusion is drawn in the last Section.

2. DEFINITION OF THE CONSTRAINTS ON MULTIPLEX PCR

The PCR constraints

In general, the primer length should be within 18 and 26 bps and the differential length of a primer pair is restricted to be smaller than 3 bps. The length constraints are defined as follows.

Let P_{pair_i} be the primer pair of forward and reverse primers, P_{2i-1} be the forward primer of a primer pair P_{pair_i}, P_{2i} be the relative reverse primer of the P_{pair_i} and a primer P is denoted as

$$P = (x_1, \ldots, x_m), \text{ where } x_k \in \{A, T, C, G \mid \forall\, k = 1, \ldots, m\}$$

$|P|$ indicates the number of nucleotides of P, which equals m. The evaluation function P_length *(P)* for the primer length constraint is defined to be

$$\text{P\_length}(P) = \begin{cases} 0, & \text{if } 18 \le |P| \le 26 \\ 1, & \text{otherwise} \end{cases}$$

Let ΔP be the difference in length of a primer pair P_{pair_i}. The evaluation function $\Delta P(P_{pair\_i})$ is defined as follows

$$\Delta \text{P}(P_{pair\_i}) = \begin{cases} 0, & \text{if } \left| |P_{2i-1}| - |P_{2i}| \right| \le 3 \\ 1, & \text{otherwise} \end{cases}$$

There is an empirical formula proposed by Wallace for calculation the melting temperature of a primer P of whose length is between 18 and 26 bps. This formula Tm(P) is written as

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

This simple formula depends directly on length and composition of the primer**.** #A indicates the amount of nucleotide "A" in P, #T indicates the amount of nucleotide "T" in P, #C and #G can be defined accordingly. Besides, the differential melting temperature of a primer pair must be under 2°C. The evaluation function ΔTm(P_{pair_i}) is denoted as

$$\Delta \text{Tm}(P_{pair\_i}) = \begin{cases} 0, & \text{if } |\text{Tm}(P_{2i-1}) - \text{Tm}(P_{2i})| \le 2 \\ 1, & \text{otherwise} \end{cases}$$

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 in certain range. In general, a proper range of GC content of a primer is from 40% to 60%. The GC content GC(P) is

$$\text{GC}(P) = \frac{\#G + \#C}{|P|} \cdot 100\%$$

where #G indicates the amount of nucleotide "G"s in P, and #C indicates the amount of nucleotide "C"s in P. The evaluation function GC_rate(P) is denoted as

$$GC\_\text{rate}(P) = \begin{cases} 0, & \text{if } 40\% \le \text{GC}(P) \le 60\% \\ 1, & \text{otherwise} \end{cases}$$

The self annealing function calculates the maximum amount of nucleotide bound to each other of two identical primers. The self-end annealing function calculates the maximum amount of nucleotide bound of a primer and forms a U turn. Figure 1 shows an example of the U form. The self annealing s_an (P) and the self-end annealing se_an (P) are defined as follows, respectively.

$$s\_\text{an}(P) = \begin{cases} 0, & \text{if } P \text{ does not cause the self - annealing} \\ 1, & \text{otherwise} \end{cases}$$

$$se\_\text{an}(P) = \begin{cases} 0, & \text{if } P \text{ does not cause the self - end annealing} \\ 1, & \text{otherwise} \end{cases}$$

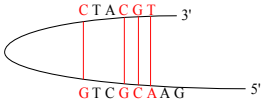

Figure 1. A primer has self-end annealing and results in the U form

The pair annealing function is similar to self annealing. The only difference is that the pair annealing calculates the maximum amount of nucleotide bound to each other of a primer pair and avoids causing the formation of primer-dimers. The pair-end annealing is similar to self-end annealing that considers binding at the 3'end between primers of a primer pair. The pair annealing P_an (P_{pair_i}) and the pair-end annealing Pe_an (P_{pair_i}) are defined as

$$\text{P\_an}(P_{pair\_i}) = \begin{cases} 0, & \text{if it does not cause primer} \\ & \text{complementary between } P_{2i-1} \text{ and } P_{2i} \\ 1, & \text{otherwise} \end{cases}$$

$$\text{Pe\_an}(P_{pair\_i}) = \begin{cases} 0, & \text{if } P_{2i-1} \text{ and } P_{2i} \text{ do not anneal} \\ & \text{at 3' end each other} \\ 1, & \text{otherwise} \end{cases}$$

The specificity of a primer is used to examine whether a unique binding position of a primer sequence within the template DNA existed.

If a primer satisfies this constraint, then it can reduce the probability of mismatch pairing and can be helpful for an efficient PCR amplification.

The evaluation function Specificity(P) is defined as

$$\text{Specificity}(P)=\begin{cases}0, \text{ if } P \text{ has the specificity in template}\\ 1, \text{ if } P \text{ does not have the specificity in template}\end{cases}$$

The GC Clamp is used to examine whether the 2 mers at the 3' end of a primer is G or C. The GC clamp of a primer GC_clamp(P) is defined as

$$\text{GC\_clamp}(P)=\begin{cases}0\ , \text{ if there are 1 - 2 base pairs of G or C at the 3' end of } P\\ 1\ , \text{ otherwise}\end{cases}$$

Constraints of multiplex PCR

In addition to each primer or primer pair satisfies the self constraints described previously and the inter-constraints among primer pairs also must be considered in the multiplex PCR reaction.

In order to make all primer pairs binding on template at the same time and prevent to produce mismatch product, the difference temperature among primer pairs need to be small in the multiplex PCR experiment. All of the differential average melting temperature of any of two primer pairs must be under 2°C. The evaluation function □Tm_mean(P_{pair_i}, P_{pair_j}) is denoted as

$$\Delta\text{Tm\_mean}(P_{pair\_i},P_{pair\_j})=\begin{cases}0\ , \text{ if } |(\frac{\text{Tm}(P_{2i-1})+\text{Tm}(P_{2i})}{2}-(\frac{\text{Tm}(P_{2j-1})+\text{Tm}(P_{2j})}{2})|\le 2\ ,\forall i\ne j\\ 1\ , \text{ otherwise}\end{cases}$$

P_{2i-1} indicates the forward primer of the primer pair P_{pair_j} in multiplex PCR, P_{2i} is the relative reverse primer of the P_{pair_i}. P_{2j-1} is the forward primer of any another primer pair P_{pair_j} in multiplex PCR, P_{2j} is the relative reverse primer of the P_{pair_j}. The individual I consists of a set of primer pairs. The evaluation function Multi_Tm(I) is defined as

$$\text{Multi\_Tm}(\ I)=\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\Delta\text{Tm\_mean}(\ P_{pair\_i},P_{pair\_j})$$

where n is the number of target sequences.

If there are n number of target sequences, then it must include 2n number of primers. By coding these 2n number of primers to serial numbers, it begins at P_1 until P_{2n} in regular order. P_{2i-1} is a forward primer in multiplex PCR. P_{2i} is a reverse primer in multiplex PCR, where i is an index from 1 to n. Figure 2 shows the conceptual graph of multiplex PCR.

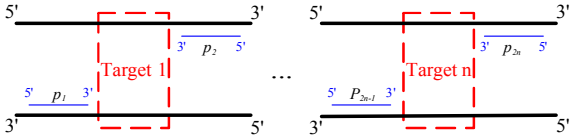

Figure 2. The conceptual graph of multiplex PCR

In multiplex PCR, the complementarity between any of two primers should be examined. Since the 3'end of a primer is more important than the 5'end of a primer, the "3'subprimer" is defined as a sub-sequence consists of the last six base pairs at 3' end of a primer. Therefore the evaluation functions are redefined as m_an (P_i, P_j) and m_e_an (P_i, P_j), respectively, as follows

$$\text{m\_an}(P_i,P_j)=\begin{cases}0\ , \text{ if it does not cause 3'subprimer complementary between } P_i \text{ and } P_j\\ 1\ , \text{ otherwise}\end{cases}$$

$$\text{m\_e\_an}(P_i,P_j)=\begin{cases}0\ , \text{ if } P_i \text{ and } P_j \text{ do not anneal at 3' end of each other}\\ 1\ , \text{ otherwise}\end{cases}$$

where P_i is a primer in multiplex PCR, P_j is the other primer in multiplex PCR.

The evaluation functions annealing Multi_an (I) and primer-end annealing Multi_e_an (I) in multiplex PCR are defined, respectively, as follows

$$\text{Multi\_an}(\ I)=\sum_{i=1}^{2n-1}\sum_{j=i+1}^{2n}\text{m\_an}(\ P_i,P_j)$$

$$\text{Multi\_e\_an}\ (I)=\sum_{i=1}^{2n-1}\sum_{j=i+1}^{2n}\text{m\_e\_an}(\ P_i,P_j)$$

where n is the number of target sequences.

The specificity is important to allow an efficient PCR amplification. However, it is very difficult to make all primers satisfy this condition in multiplex PCR. Instead to satisfy the strict specificity constraint in PCR experiment, one can allow a less restricted condition - Secondly specificity to assist the specificity constraint. The secondary specificity pays attention to examine the 3'subprimer of a primer, when a primer P cannot satisfy the specificity constraint at a position of the template DNA, but the 3'subprimer of P does not complement at the same position of the template DNA. An example is shown in Figure 3. The blue color of the sequence is the 3'subprimer of P. It is obvious that this primer does not satisfy the specificity constraint. However the 3'subprimer of P does not bind with the template T. Hence, P satisfies the secondly specificity constraint. The evaluation function SSpecificity(P) is defined as

$$\text{SSpecificity}(P)=\begin{cases}0, \text{if Specificity}(P)=0\\ 0, \text{if Specificity}(P)\ne 0 \text{ and 3'subprimer of } P \text{ appears in template once}\\ 1, \text{otherwise}\end{cases}$$

Figure 3. An example for a primer satisfied the secondly specificity constraint

The evaluation function Multi_SSpecificity(I) in multiplex PCR is defined as

$$\text{Multi\_SSpecificity}(I) = \sum_{i=1}^{2n} \text{SSpecificity}(P_i)$$

When there are more than one target DNA sequences in one multiplex PCR experiment, any of two target sequence lengths that should differ over certain value in order to easily distinguish lengths of products in the electrophoresis. In general, the differential value is restricted within 50-100. The evaluation function □T_length(*I*) of the electrical mobility is denoted as

$$\Delta\text{T\_length}(I) = \begin{cases} 0\ , \text{ if } 50 \le \left|\left|T\text{arget}_{i+1}\right| - \left|T\text{arget}_i\right|\right| \le 100 \\ \qquad \forall\ 1 \le i < n \\ 1\ , \text{ otherwise} \end{cases}$$

where |Target$_i$| indicates one of the target length in multiplex PCR .

3. The proposed algorithm

The goal of the proposed algorithm is to design the primer pairs for multiplex PCR reaction. The proposed algorithm consists of initialization process, evaluation process, selection process, crossover process and mutation process. Figures 4 and 5 illustrate these processes of the proposed algorithm. Before performing the process, it is very important to find an appropriate coding namely the qualified chromosome for the proposed algorithm. Since there are many primer pairs in a multiplex PCR experiment, in order to express the multiplex PCR primer pairs, it is transformed into the independent form and is suitable for the GA operations. The independent form is denoted as

$$(F_1, \alpha_1, \beta_1, \gamma_1, F_2, \alpha_2, \beta_2, \gamma_2, ..., F_n, \alpha_n, \beta_n, \gamma_n)$$

where "*F*" is the start position of the forward primer of the individual toward the template DNA sequence. "*α"* denotes the length of the forward primer toward the *Target*. "*β"* indicates the length of the *Target*. "*γ"* denotes the length of the reverse primer toward the *Target*. "n" is the amount of target sequences. *α* and *γ* must obey the length constraint of a primer. The total sum of *α*, *β* and *γ* must be larger than the length of duplicate DNA sequence, which is assigned by the user. However, one has to transform the independent form back to dependent form in order to calculate the fitness value. The dependent form for the *Target* $_n$ is calculated as follows:

$$\begin{aligned} F_s^n &= F_n \\ F_e^n &= F_n + \alpha_n\text{-}1 \\ R_s^n &= F_n + \alpha_n + \beta_n \\ R_e^n &= F_n + \alpha_n + \beta_n + \gamma_n - 1 \end{aligned}$$

where F_s^n is the beginning position of a forward primer F_e^n is the end position of a forward primer; R_s^n is the beginning position of a reverse primer; R_e^n is the end position of a reverse primer. Based on these values, we can express the clamping region of the *Target* $_n$

The initialization step is to produce a quantity of individuals randomly and compose these individuals to be the population for the proposed algorithm.

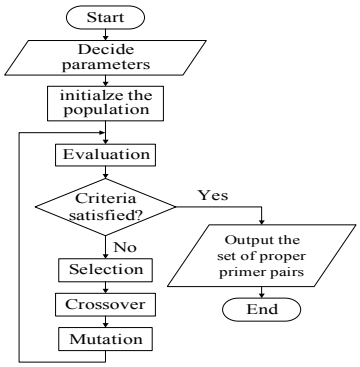

Figure 4. The flowchart of the proposed algorithm

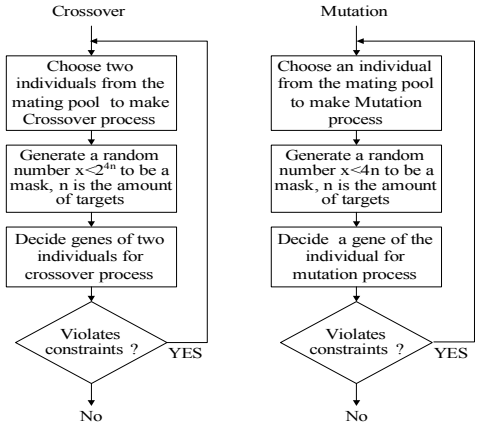

Figure 5. T he flowchart of the crossover process and the mutation process

The initialization step is to produce a quantity of individuals randomly and compose these individuals to be the population for the proposed algorithm. The genes of each individual all must satisfy the constraints in below. If an individual does not obey one of these constraints, then it must be regenerated. Performing the initialization step, until all individuals follow the constraints and the amount of individuals is equal to the population size.

All the constraints of multiplex PCR primer design mentioned before are combined into an objective function fitness(*I*) as follows.

$$\begin{aligned} \text{fitness}(I) = {} & \text{P\_length}(I) + \Delta\text{P}(I) + \Delta\text{T\_length}\ (I) \\ & + \text{GC\_rate}(I) + 2 * \text{GC\_clamp}(I)\ ) \\ & + (\text{s\_an}(I) + \text{se\_an}(I) + \text{multi\_an}(I) \\ & + \text{multi\_e\_an}(I)) \\ & + 30 * (\text{Multi\_SSpecificity}(I) + \Delta\text{Tm}(I) \\ & + \text{Multi\_Tm}(I)) \end{aligned}$$

where *I* is the generic of all individuals (all sets of primer pairs). Each individual is evaluated within the fitness function and can obtain a fitness value for each individual. When this fitness value is smaller, the primer pairs of this individual are more suitable for the multiplex PCR experiment. Each constraint is assigned by a suitable penalty value that is obtained by applying the orthogonal arrays method.

The selection process is based on the way of the Roulette Wheel Selection method by using the fitness value in the evaluation step to decide the probability of each individual to be preserved to the next generation. The duplicate individual is sent into the mating pool and waited for the operation of the crossover process and the mutation process.

The crossover process is operated on two duplicate individuals that are sent into the mating pool. Before performing the crossover process, a random number X must be generated. The number X must be smaller than 2^{4n} where n is the amount of target sequences. It is used as a mask to transform X into the 4n bits of binary notation. This mask is used to decide the genes of two individuals to perform the crossover process. Figure 6 shows the concept of the crossover process of the proposed algorithm.

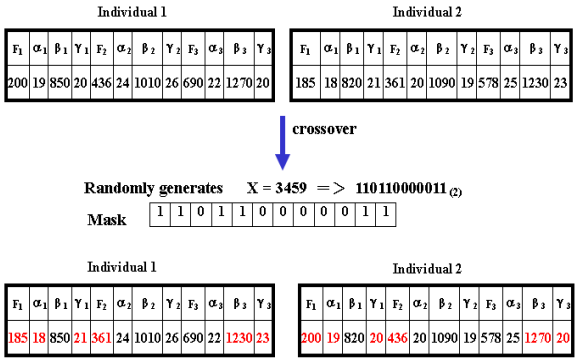

Figure 6. The concept of the crossover process

The mutation process is similar to the crossover process, which is to select one duplicate individual into the mating pool and decide whether this individual perform the mutation process according to the mutation rate. Before performing the mutation process, a random number Y must be generated. The number Y must be smaller than 4n, where n is the target amount. It uses this number Y as a mask. By using the mask, the chosen gene of an individual is changed and further checking the individual satisfied the length constraint or not. Figure 7 shows the concept of the mutation process of the proposed algorithm.

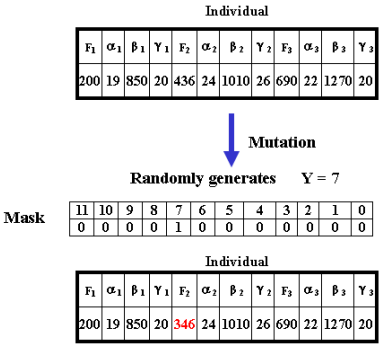

Figure 7. The concept of the mutation process

In order to efficiently calculate the specificity, we propose a pattern model, called MAtch Pattern model (MAP-model).

Let S' be the complementary sequence array of a template sequence S, and it is defined as

$$S = (s_1, \ldots, s_n), \text{ where } s_i \in \{A,T,C,G \mid \forall\, i = 1, \ldots, n\}$$

$$S' = [s'_i],\ s'_i = \begin{cases} A, \text{ when } s_i = T \\ C, \text{ when } s_i = G \\ G, \text{ when } s_i = C \\ T, \text{ when } s_i = A \end{cases},\ \forall\ 1 \le i \le n$$

P is the sequence array of a primer, and it is denoted as

$$P = [p_i], \text{where} p_i \subseteq \{A,T,C,G\},\ p_i \neq \phi, \forall 1 \le i \le m$$

where n and m are the lengths of S' and P. We use four arrays to record the composition of four kinds of nucleotides in S'. The size of each pattern equals n. These four patterns are defined as follows

$$S_A = [a_i], \text{where} a_i = \begin{cases} 0, \text{otherwise} \\ 1, \text{if } s'_i \text{ is the nucleotide "A"} \end{cases},\ \forall 1 \le i \le n$$

$$S_T = [t_i], \text{where } t_i = \begin{cases} 0, \text{otherwise} \\ 1, \text{if } s'_i \text{ is the nucleotide "T"} \end{cases},\ \forall 1 \le i \le n$$

$$S_C = [c_i], \text{where } c_i = \begin{cases} 0, \text{otherwise} \\ 1, \text{if } s'_i \text{ is the nucleotide "C"} \end{cases},\ \forall 1 \le i \le n$$

$$S_G = [g_i], \text{where } g_i = \begin{cases} 0, \text{otherwise} \\ 1, \text{if } s'_i \text{ is the nucleotide "G"} \end{cases},\ \forall 1 \le i \le n$$

Define a matrix X to examine the specificity of a primer with a template sequence. X depends on the composition of P.

$$X = [x_{ij}], \text{where } x_{ij} = \begin{cases} a_j, \text{if } p_i \text{ is the nucleotide "A"} \\ t_j, \text{if } p_i \text{ is the nucleotide "T"} \\ c_j, \text{if } p_i \text{ is the nucleotide "C"} \\ g_j, \text{if } p_i \text{ is the nucleotide "G"} \end{cases} \quad \forall 1 \le i \le m,\ 1 \le j \le n$$

For example, P = [A T A A C C G G A T A]. Then X^T is

$$X^T = \{S_A^T, S_T^T, S_A^T, S_A^T, S_C^T, S_C^T, S_G^T, S_G^T, S_A^T, S_T^T, S_A^T\}$$

where X^T is the transpose of X. It computes sum of all entries in the same diagonal direction. The sum means the amount of nucleotides of the primer binding to the template. The algorithm searches the maximum amount of binding other than the correct binding site which the maximum probability of mis-binding, it checks the percentage of bindings to determine whether the primer satisfies the specificity or not. The function $M_{bind}(X)$ is used to compute the maximum amount of nucleotides that a primer is bound with a template sequence except the correct binding site and it is defined as

$$M_{bind}(X) = \max_{k=0,\ldots,n-m} \sum_{i=1}^{m} x_{ij}, \forall\ j = i + k$$

An example is given in the following to explain the MAP model. A primer is "CTAGC", and a template S is "CTTAGACCG". Thus P = [C T A G C], and S' = [G A A T C T G G C]. The length of P equals 5. The length of S' equals 9. According to the four kinds of nucleotides in the S', the array S_A=[0 1 1 0 0 0 0 0 0], the array S_T =[0 0 0 1 0 1 0 0 0], the array S_C=[0 0 0 0 1 0 0 0 1] and the array S_G =[1 0 0 0 0 0 1 1 0]. Then, it maps each nucleotide of P to the same nucleotide pattern in order. The reference matrix X is used to examine the specificity. By evaluating the function $M_{bind}(X)$, the maximum amount of nucleotides that a primer bound with a template sequence can be obtained. All cases of the primer binding with the template are shown in Figure 8. The maximum amount of nucleotides is $M_{bind}(X)$ = max(2,0,1,1,4) = 4. It means that there are most four nucleotides binding on the template in this case. By using the MAP model, it is able to examine the specificity faster and easier.

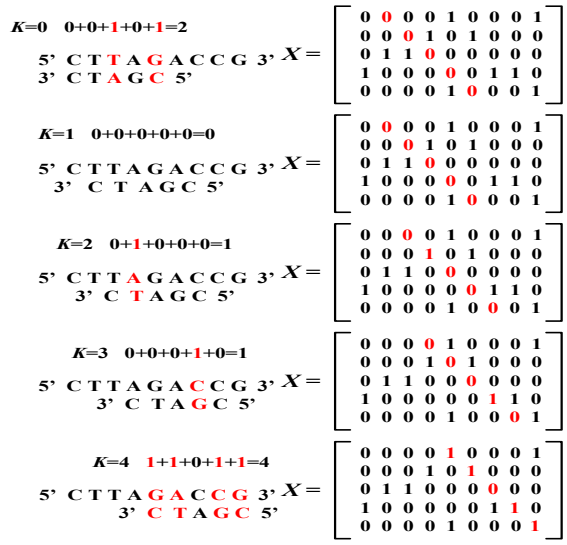

Figure 8. All cases of *P* binding with the *S*

Most of existing software or researches ignore the specificity constraint. The adapted method for checking the specificity constraint is just to align a nucleotide after a nucleotide between a primer and a template. However, this way takes too much time, which spends almost the same time as all other constraints do. The proposed algorithm uses the MAP model to reduce the running time. Although it still spends lots of time to handle the specificity, it reduces much more time than the original method. These two different ways are compared in Table 1. Assume there are m template sequences, the average length of m number of template sequences is n. According to the amount of templates, it needs twice amount of primers to amplify the sequences and the average length of 2m number of primers is P. In Table 1, the total number of the operation of comparing between primers and template sequences are calculated. For one base pair, the average comparison is 1*1/4+2*1/4+3*1/4+4*1/4=10/4=2.5. Hence, for a primer with length P, total comparison times are 2.5P times. When using the traditional method to evaluate the specificity, it needs to check (n-P+1) positions for a template sequence. When using the MAP model to evaluate the specificity, it just considers the composition of a primer to produce matrix X instead. However, the total numbers of the operation on additions for these two methods are the same. As a result, the proposed algorithm not only find primer pairs satisfy all the constraints but also more efficient in solving the problem.

Table 1. The total number of the operation of comparing by using the traditional method and the MAP model

The number of average comparing		
	Traditional method	MAP-model
one base pair of a primer	2.5	2.5
a primer	2.5P	2.5P
(n-P+1) positions	2.5P*(n-P+1)	
m templates	2.5P*(n-P+1)*m	2.5*P*m
2m primers	2.5P*(n-P+1)*m*2m	2.5*P*m*2m

4. Dry Dock Experiments

4.1 cDNA templates

Five gene family cDNA templates are used as our experimental DNA templates, Homo sapiens annexin A1 (ANXA1), mRNA (GenBank Acc#NM_000700), Homo sapiens annexin A2 (ANXA2), transcript variant 3, mRNA(GenBank Acc#NM_004039), Homo sapiens annexin A3 (ANXA3), mRNA (GenBank Acc#NM_005139), Homo sapiens annexin A5 (ANXA5), mRNA (GenBank Acc#NM_001154) and Homo sapiens annexin A13 (ANXA13), mRNA (GenBank Acc#NM_004306) [10].

4.2 The environment

The parameter settings are given as follows. The population size is 1000, the crossover rate is 80% and the mutation rate is starting 1%. With the generation becoming large, the mutation rate can grow up to 90%. The proposed algorithm is implemented on Intel Celeron 1G Hz, 128 MB, Windows 2000 and written in BCB.6.0.

4.3 Results

The results of dry dock experiments are listed in Tables 2-4. In Table 2, three gene family cDNA sequences ANXA1, ANXA2 and ANXA3 are used for the experiment,. The multiplex product sizes of the solution obtained from the proposed algorithm are 185 bps, 270 bps and 379 bps. Three similar product sizes are selected based on the algorithm proposed by Schoske which is implemented from the other research by using the Primer3. Experiments show that the primer pairs found by the proposed algorithm meet all of the constraints. Furthermore, all of primers have the specificity which is useful for amplifying gene family sequences respectively. In contrast, the primer pairs found by Primer3 have barely satisfied the specificity. However, the primer pairs found by Primer3 also satisfied most of the constraints, but do not satisfy the GC clamp completely. Four gene family cDNA sequences used for the second experiment are ANXA1, ANXA2, ANXA3 and ANXA5. There are four primer pairs of different product sizes obtained by the proposed algorithm in Table 3. The melting temperatures of them are all within suitable experimental temperature and all of them satisfies the GC clamp. It is more significant that all of them also have the specificity. In Table 4, we try to select more primer pairs of different product sizes. Five gene family cDNA sequences used for third experiment are ANXA1, ANXA2, ANXA3, ANXA5 and ANXA13. It shows all primer pairs also satisfy all of the constraints including the complementary and the specificity. Furthermore, the temperature differences among all primer pairs are small and all within suitable experimental temperature. Although the electrical mobility are not satisfied completely, it is still within acceptable range. Other constraints of these primer pairs are all satisfied and feasible for performing the PCR experiment.

4.4 Discussions

Multiplex PCR primer design is more difficult than a single PCR primer design, since it considers not only the constraints of single primer or single primer pairs but also needs to consider the constraints among different primer pairs for different target sequences.

Table 2. The comparison of primer pairs found by the proposed algorithm and Primer3

	The proposed algorithm			**Primer3**		
Target size	150 bps	250 bps	350 bps	150 bps	250 bps	350 bps
Forward primer (5'->3')	**AGACTTGGCT GATTCAGATG**	**GTAGAAGAGC AGAGGATGG**	**GCGGCAGCTG ATTGTTAAG**	**AAGGTGTGGA TGAAGCAACC**	**CTCTACACCCC CAAGTGCAT**	**ATGGCATCTAT CTGGGTTGG**
Reverse primer (5'->3')	**CTTTCAACTCC AGGTCCAG**	**GCTTGTTCTGA ATGCACTG**	**ATCTTGTTTGG CCAGATGC**	**ATTGCGCTGGA GTTTTTAGC**	**CAAAATCACC GTCTCCAGGT**	**ATCCTTCATTT GCCTGCTTG**
Position (F)	**661-680**	**578-596**	**201-219**	**247-266**	**100-119**	**47-66**
Position (R)	**827-845**	**829-847**	**561-579**	**426-445**	**349-368**	**427-446**
Product size (bp)	**185**	**270**	**379**	**180**	**250**	**381**
Primer length (F/R) (mer)	**20/19**	**19/19**	**19/19**	**20/20**	**20/20**	**20/20**
Melting temperature (F/R) (^{o}C)	**58/58**	**58/56**	**58/56**	**60/58**	**62/60**	**60/58**
Temperature difference (^{o}C)	**0**	**2**	**2**	**2**	**2**	**2**
GC-content (F) (%)	**45**	**53**	**53**	**50**	**55**	**50**
GC-content (R) (%)	**53**	**47**	**47**	**45**	**50**	**45**
GC clamp (F)	**G**	**GG**	**G**	**CC**	**T**	**GG**
GC clamp (R)	**G**	**G**	**GC**	**GC**	**T**	**G**
Self-annealing	**No**	**No**	**No**	**No**	**No**	**No**
Pair-annealing	**No**	**No**	**No**	**No**	**No**	**No**
Specificity	**Yes**	**Yes**	**Yes**	**No**	**No**	**No**

Table 3. Four primer pairs found by the proposed algorithm

Target size	100 bps	200 bps	300 bps	400 bps
Forward primer (5'->3')	**GTCAACAGATCAAAGC AGC**	**ACAGCCATCAAGA CCAAAG**	**AGACTTACTGTTG GCCATAG**	**ACTGACTTCCCT GGATTTG**
Reverse primer (5'->3')	**GCATCAAATTGCGCTG GAG**	**AAGCGTCATACTG AGCAGG**	**GAGAAGAAGTAA GGTGGAG**	**CTGGTAGTACCC TGAAGTG**
Position (F)	**303-321**	**178-196**	**750-769**	**219-237**
Position (R)	**414-432**	**364-382**	**1058-1082**	**623-641**
Product size (bp)	**130**	**205**	**327**	**423**
Primer length (F/R) (mer)	**19/19**	**19/19**	**20/19**	**19/19**
Melting temperature (F/R) (^{o}C)	**56/58**	**56/58**	**58/56**	**56/58**
Temperature difference (^{o}C)	**2**	**2**	**2**	**2**
GC-content (F) (%)	**47**	**47**	**45**	**47**
GC-content (R) (%)	**53**	**53**	**47**	**53**
GC clamp (F)	**GC**	**G**	**G**	**G**
GC clamp (R)	**G**	**GG**	**G**	**G**
Self-annealing	**No**	**No**	**No**	**No**
Pair-annealing	**No**	**No**	**No**	**No**
Specificity	**Yes**	**Yes**	**Yes**	**Yes**

Table 4. Five primer pairs found by the proposed algorithm

Target siz	100 bps	150 bps	200 bps	250 bps	300 bps
Forward primer (5'->3')	**GACAGACGTAAACGTGTTC**	**GTAGAAGAGCAGAGGATGG**	**TAGTGACTCCACCAGCAG**	**GTTTGGCAGGGATCTTCTG**	**AAAGATCTAGCTGGTCAGG**
Reverse primer (5'->3')	**CTTTCAACTCCAGGTCCAG**	**GGGCTGTAACTCTTGTACC**	**CATCTGCCAAAGTCAACAG**	**AACATCCGCTGGTAGTACC**	**CACGACTATGCGAATCAAC**
Position (F)	**718-736**	**578-596**	**305-322**	**371-389**	**580-598**
Position (R)	**827-845**	**740-758**	**511-529**	**631-649**	**879-897**
Product size (bp)	**128**	**181**	**225**	**279**	**318**
Primer length (F/R) (mer)	**19/19**	**19/19**	**18/19**	**19/19**	**19/19**
Melting temperature (F/R) (^{o}C)	**56/58**	**58/58**	**56/56**	**58/58**	**56/56**
Temperature difference (^{o}C)	**2**	**0**	**0**	**0**	**0**
GC-content (F) (%)	**47**	**53**	**56**	**53**	**47**
GC-content (R) (%)	**53**	**53**	**47**	**53**	**47**
GC clamp (F)	**C**	**GG**	**G**	**G**	**GG**
GC clamp (R)	**G**	**CC**	**G**	**CC**	**C**
Self-annealing	**No**	**No**	**No**	**No**	**No**
Pair-annealing	**No**	**No**	**No**	**No**	**No**
Specificity	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**

Though the proposed algorithm uses the MAP model to reduce the execution time of the specificity, the specificity checking still takes a major proportion of total execution. Without considering the specificity and the correlation among primers, the proposed algorithm obtains the similar quality of dry-duck experimental result as Primer 3 for single primer design.

5. Conclusions

Multiplex PCR primer design is a complex problem that has several constraints on selecting primer pairs need to be satisfied for a successful PCR experiment. In this paper, we have proposed the new method for the multiplex PCR experiment. We also proposed a MAP model to reduce the computing time, when solve the specificity constraint that allows many primer pairs to bind with a unique position within the template DNA and do not interact to each other. The dry-dock experiment shows that the proposed algorithm has found a set of suitable primer pairs of gene family sequences for the multiplex PCR experiment in a single run. This set of primer pairs that has satisfied most constraints mentioned above and can clip out multiple targets of gene family cDNA sequences respectively by satisfing the specificity constraint.

6. ACKNOWLEDGMENTS

Our thanks to Yow-Ling Shiue, the professor of Institute of Biomedical Sciences, National Sun Yat-Sen University, Kaohsiung, Taiwan, for offering advices to all the design constraints of the Multiplex PCR primer design problem.

7. REFERENCES

[1] McPherson, M. J., Quirke, P. and Taylor, G. R. *PCR: A Practical Approach.* Oxford University Press, New York, 1993

[2] Singh, V. K., Mangalam, A. K., Dwived, S. and Naik, S. Primer premier : Program for design of degenerate primers form a protein sequence. *BioTechniques*, 24, 1998, 318-319.

[3] Lincoln, S. E., Daly, M. J. and Lander, E. S. PRIMER: A Computer Program for Automatically Selecting PCR Primers. MIT Center for Genome Research and Whitehead Institute for Biomedical Research, Cambridge, Massachusetts, 1991.

[4] Nicodème, P. A computer support for genotyping by multiplex PCR. Technical Report LIX/RR/93/09, LIX, Ecole Polytechnique, 91128, Palaiseau Cedex, France, 1993.

[5] Nicodème, P. and Steyaert, J.. (1997) Selecting optimal oligonucleotide primers for multiplex PCR. *Intelligent Syst. Mol. Biol*. Vol. 51, pp. 210-213.

[6] Schoske, R. , Vallone, P. M., Ruitberg, C. M. and Butler, J. M. Multiplex PCR design strategy used for the simultaneous amplification of 10 Y chromosome short tandem repeat (STR) loci. *Anal. Bioanal. Chem.*, 375, 2003, 333-343.

[7] Kämpke, T., Kieninger, M. and Mecklenburg, M. Efficient primer design algorithms, *Bioinformatics*, 17, 2001, 214-225.

[8] Goldberg, D. E. *Genetic Algorithms in Search, Optimization, and Machine Learning.* Addison-Wesley, New York, 1989.

[9] Wu, J. S., Lee, C. N., Wu, C. C. and Shiue, Y. L. Primer Design Using Genetic Algorithm, *Bioinformatics*, 20, 2004, pp. 1710-1717.

[10] Arbour, N. C., Lorenz, E., Schutte, B. C., Zabner, J., Kline, J. N., Jones, M., Frees, K., Watt, J. L., and Schwartz, D. A. A genetics basis for a blunted response to endotoxin in Humans, 2000. *http://www.ncbi.nlm.nih.gov/entrez/.*