Mutagenic Primer Design for Mismatch PCR-RFLP SNP Genotyping Using a Genetic Algorithm

Cheng-Hong Yang, Yu-Huei Cheng, Cheng-Huei Yang, and Li-Yeh Chuang

Abstract—Polymerase chain reaction-restriction fragment length polymorphism (PCR-RFLP) is useful in small-scale basic research studies of complex genetic diseases that are associated with single nucleotide polymorphism (SNP). Designing a feasible primer pair is an important work before performing PCR-RFLP for SNP genotyping. However, in many cases, restriction enzymes to discriminate the target SNP resulting in the primer design is not applicable. A mutagenic primer is introduced to solve this problem. GA-based Mismatch PCR-RFLP Primers Design (GAMPD) provides a method that uses a genetic algorithm to search for optimal mutagenic primers and available restriction enzymes from REBASE. In order to improve the efficiency of the proposed method, a mutagenic matrix is employed to judge whether a hypothetical mutagenic primer can discriminate the target SNP by digestion with available restriction enzymes. The available restriction enzymes for the target SNP are mined by the updated core of SNP-RFLPing. GAMPD has been used to simulate the SNPs in the human SLC6A4 gene under different parameter settings and compared with SNP Cutter for mismatch PCR-RFLP primer design. The *in silico* simulation of the proposed GAMPD program showed that it designs mismatch PCR-RFLP primers. The GAMPD program is implemented in JAVA and is freely available at http://bio.kuas.edu.tw/gampd/.

Index Terms—Polymerase chain reaction (PCR), restriction fragment length polymorphism (RFLP), single nucleotide polymorphism (SNP), mutagenic primer design, genetic algorithm (GA).

1 Introduction

Polymerase chain reaction-restriction fragment length polymorphism (PCR-RFLP) is a simple, inexpensive, accurate, and common laboratory technique used to gain an knowledge of the causes of genetic variants and mutations; it is especially useful in small basic research studies of complex genetic diseases [1], [2]. It is frequently used for the detection of single nucleotide polymorphisms (SNPs). SNPs are the most common genetic variants and play an important role in population genetics and evolutionary studies [3], pharmacogenetic analysis [4], malignancy studies [5], [6], preventive medicine [7], [8], personalized medicine [9], and forensics [10]. However, many SNPs cannot be genotyped by PCR-RFLP, because a restriction enzyme to discriminate the SNPs by digestion does not exist. This "mismatch PCR-RFLP" primer design problem is of great importance and has to be addressed.

There are currently only a few systems that provide mismatch PCR-RFLP primer design. For example, V-MitoSNP identifies available restriction enzymes from REBASE [11] and designs a PCR-RFLP primer set for RFLPs in all mtSNPs [12]. However, this system uses very simple constraints and generally identifies primers lacking in stringent quality; it can only design primers for mitochondrial SNPs. SNP Cutter uses a preselected or customizable list of restriction enzymes and uses Primer3, the most popular noncommercial primer design software [13], [14], [15], to look for PCR-RFLP primer sets. Many SNP IDs are not found, and it does not provide the information pertaining to the latest restriction enzymes. It thus often omits feasible solutions. Prim-SNPing is an improved software tool with a mismatch PCR-RFLP primer design function for cost-effective SNP genotyping [16]. However, the incorporated window-sliding strategy limits its search efficiency and the quality. Therefore, the development of an improved method for mismatch PCR-RFLP primer design and software based on it is still mandated.

Development of a mismatch PCR-RFLP primer design method is a challenging task as numerous primer constraints must be conformed to, e.g., primer length, length difference, melting temperature (T_m), T_m difference, GC proportion, GC clamp, dimer (including cross-dimer and self-dimer), hairpin structure, specificity, and PCR product size [13], [17], [18], [19], [20], [21]. A primer furthermore must contain additional mutagenic base(s) adjacent to the target SNP site [22], [23] which allow the target SNP to be discriminated by digestion with available restriction enzymes.

- *C.-H. Yang is with the Department of Network Systems, Toko University, Chiayi, Taiwan and the Department of Electronic Engineering, National Kaohsiung University of Applied Sciences, Kaohsiung, Taiwan. E-mail: chyang@cc.kuas.edu.tw.*
- *Y.-H. Cheng is with the Department of Network Systems, Toko University, Chiayi, Taiwan. E-mail: yuhuei.cheng@gmail.com.*
- *C.-H. Yang is with the Department of Electronic Communication Engineering, National Kaohsiung Marine University, No. 142, Hai-chuan Rd., Nan-tzu, Kaohsiung 81157, Taiwan. E-mail: chyang@mail.nkmu.edu.tw.*
- *L.-Y. Chuang is with the Department of Chemical Engineering and Institute of Biotechnology and Chemical Engineering, I-Shou University, No. 1, Sec. 1, Syuecheng Rd., Dashu District, Kaohsiung 84001, Taiwan. E-mail: chuang@isu.edu.tw.*

Manuscript received 28 Feb. 2011; revised 11 Sept. 2011; accepted 29 Sept. 2011; published online 30 Jan. 2012.
For information on obtaining reprints of this article, please send e-mail to: tcbb@computer.org, and reference IEEECS Log Number TCBB-2011-02-0048.
Digital Object Identifier no. 10.1109/TCBB.2012.25.

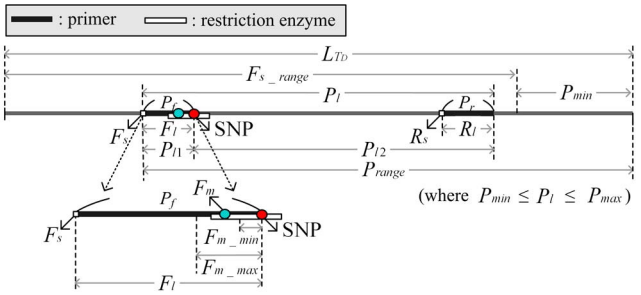

Fig. 1. Diagram for designing mismatch PCR-RFLP primers. The length of the template DNA is L_{TD}, the minimum PCR product length is P_{min}, the maximum PCR product length is P_{max}, the length of the forward primer is F_l, the mutagenic position of the forward primer is F_m, the PCR product length between the start position of forward primer and the end position of reverse primer is P_l, the length of the reverse primer is R_l, the PCR product length between the start position of forward primer and SNP is P_{l1}, the PCR product length between the SNP and the end position of reverse primer is P_{l2}.

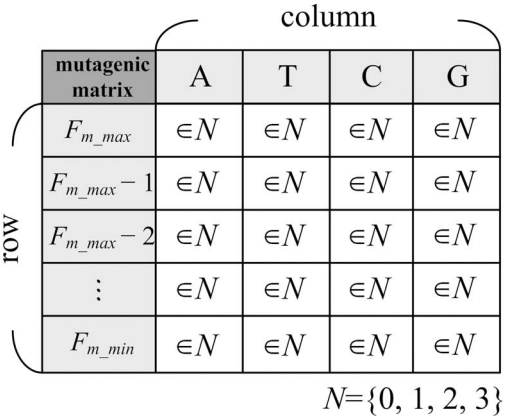

Fig. 2. Diagram of the mutagenic matrix. This mutagenic matrix contains $(F_{m\_max} - F_{m\_min} + 1) \times 4$ elements. The row field represents the mutagenic nucleotide candidates, which are composed by nucleotides between $(SNP - F_{m\_max} + 1)$ and $(SNP - F_{m\_min} + 1)$ in T_D. The column field represents the four mutagenic nucleotides "A," "T," "C," and "G." In the mutagenic matrix, the internal element values are 0, 1, 2, or 3, which represent different RFLP meanings.

In this study, we introduce a genetic algorithm (GA) [24], [25] and use the updated core of SNP-RFLPing [26], [27], [28] to provide the design for mismatch PCR-RFLP primer sets. GA is a well-known stochastic search algorithm modeled on the process of natural selection [29]. It constitutes the optimization technique underlying biological evolution. In a GA, the evolutionary computations involved, i.e., selection, crossover, mutation, and replacement, are effective in searching for optimal solutions for many mutagenic primer sets. After each run, individuals in a GA share information with each other. The superior solutions based on a fitness rule are refined from generation to generation. SNP-RFLPing is a time-saving application for mining the restriction enzymes for RFLP assays. The core of SNP-RFLPing has recently been updated and now provides more robust results from REBASE. Consequently, feasible mismatch PCR-RFLP primer pairs are mined.

2 Materials and Methods

2.1 SNP Template Sequences

A point mutation in the SLC6A4 gene was recently identified and shown to be associated with autism spectrum disorders [30], psychosis [31], and bipolarity [32] in patients. The SLC6A4 gene is the member 4 for the solute carrier family 6 (neurotransmitter transporters, serotonin). A total of 288 SNPs in the SLC6A4 gene are used as template sequences. Deletion/insertion polymorphism (DIP) and multinucleotide polymorphism (MNP) were excluded. Each template sequence has a flanking length of 500 nt and was retrieved by SNP-Flankplus (http://bio.kuas.edu.tw/snp-flankplus/) [33]. The reference cluster IDs of the SNPs are shown in http://bio.kuas.edu.tw/gampd/dataset.jsp.

2.2 Problem Definition

Let T_D be the template DNA sequence made up of nucleotides "A," "T," "C," or "G," of the DNA with an identified SNP. The identified SNP, i.e., the target SNP, is represented by the IUPAC code (M, R, W, S, Y, K, V, H, D, B, or N) or the dNTP format ([dNTP1/dNTP2]). In this study, we focused only on the true SNPs as described in dbSNP [34] of NCBI, i.e., deletion/insertion polymorphism (DIP or Indel) and multinucleotide polymorphisms are not included. A forward primer P_f with a length F_l, i.e., a mutagenic primer, includes the target SNP and a mutagenic base adjacent to the target SNP site. We define F_m as the mutagenic position which lies between F_{m_min} and F_{m_max} starting from the SNP site. The mutagenic primer can assist the target SNP discriminated by digestion with one available restriction enzyme at least. A reverse primer P_r with a length R_l makes up a primer pair with the forward primer P_f is a complementary and reverse subsequence in T_D without any mutagenic positions. The PCR product length between P_f and P_r is P_l. Now, we define a vector $P_v = (F_m, F_1, P_1, R_1, \Sigma)$ to represent a solution for the mismatch PCR-RFLP primer pair. The symbol Σ is the integer 0, 1, 2, or 3 and represents mutagenic nucleotide "A," "T," "C," or "G," respectively. All used parameters are described in Fig. 1.

2.3 Mutagenic Matrix for Selecting the Mutagenic Primer

A mutagenic matrix is generated to judge whether a hypothetical mutagenic nucleotide allows the target SNP to be discriminated by digestion with available restriction enzymes. This mutagenic matrix contains $(F_{m\_max} - F_{m\_min} + 1) \times 4$ elements (Fig. 2). The row fields represent the mutagenic position lying between $(SNP - F_{m\_max} + 1)$ and $(SNP - F_{m\_min} + 1)$ in T_D. The column fields represent the mutagenic nucleotides "A," "T," "C and "G." In the mutagenic matrix, the internal element values are 0, 1, 2, or 3, which represent different RFLP meanings. The value 0 means that when a mutagenic nucleotide is determined, both the sense and antisense strand have available restriction enzymes to discriminate the target SNP. The value 1 means that when a mutagenic nucleotide is determined, either the sense or antisense strand has an available restriction enzyme to discriminate the target SNP. The value 2 means that when a mutagenic nucleotide is determined, neither the sense nor antisense strand has an available restriction enzyme to discriminate the target SNP. And finally, the value 3 means that no mutagenic nucleotide is designed; the original nucleotide is retained.

2.4 Primer Constraints

In order to design a pair of good mismatch PCR-RFLP primers, primer constraints are important. Here, we consider the primer constraints primer length, length difference, melting temperature (T_m), T_m difference, GC proportion, GC clamp, dimer (including cross-dimer and self-dimer), hairpin structure, specificity and PCR product size [13], [17], [18], [19], [20], [21]. They influence the efficiency of the PCR amplification. The primer length constraint avoids primers that are too long or too short. An appropriate primer length leads to very specific sequences. A typical primer length is between 16 and 28 nt. In a PCR experiment, a length difference no more than 3 nt between the forward primer and the reverse primer is considered optimal [20]. The T_m of a primer should be between $45°\mathrm{C}$ and $62°\mathrm{C}$ and the T_m difference should not exceed $5°\mathrm{C}$. An unsuitable T_m difference can lead to the amplification of nonspecific products. The GC proportion is a value indicating the ratio of the nucleotides "G" and "C" in a primer. An appropriate GC proportion provides a sufficient thermal window for efficient annealing. In general, this proportion lies between 40 and 60 percent. GC clamp makes sure that the $3'$ terminal end of a primer is nucleotide "G" or "C" and ensures a tight localized hybridization bond [13]. The occurrence of dimers during a PCR experiment is problematic. Dimers include cross-dimers and self-dimers. A cross-dimer is formed when P_f and P_r anneal to each other, and a self-dimer is formed when P_f and P_f or P_r and P_r anneal to each other. Consequently, dimers should be avoided. A hairpin occurs when a primer anneals to itself. Eliminating such an annealed primer is one of the objectives of primer design. Furthermore, the specificity ensures that the designed P_f and P_r primers do not reappear in T_D and thus raise the success rate of PCR experiments. The PCR product size must be greater than 100 nt in order to distinguish PCR bands via gel electrophoresis.

2.5 Method for Designing Mismatch PCR Primers

Six processes are applied in the proposed method, a flowchart of which is shown in Fig. 3. The processes are

1. calculation of the mutagenic matrix;
2. judgment on the existence of mutagenic primers;
3. creation of a random initial primer set;
4. evaluation of a primer pair score;
5. judgment of termination criteria, and
6. operations of selection, crossover, mutation and replacement, respectively.

These processes are described below:

1. *Calculation of mutagenic matrix.* First, the mutagenic matrix is generated based on the nucleotides adjacent to the target SNP. This saves time for the fitness estimation in the proposed method.

2. *Judgment on the existence of mutagenic primers.* The proposed method judges whether mutagenic primers exist (i.e., the element values 0 or 1 are in the "mutagenic matrix"). The target SNP must be recognized by the restriction enzymes when the mutagenic nucleotide is determined. When the mutagenic matrix has no any available restriction enzymes to distinguish the allele of the target SNP (i.e., the element values are 2 or 3 in the "mutagenic matrix"), the mismatch PCR-RFLP primer design will not be significant.

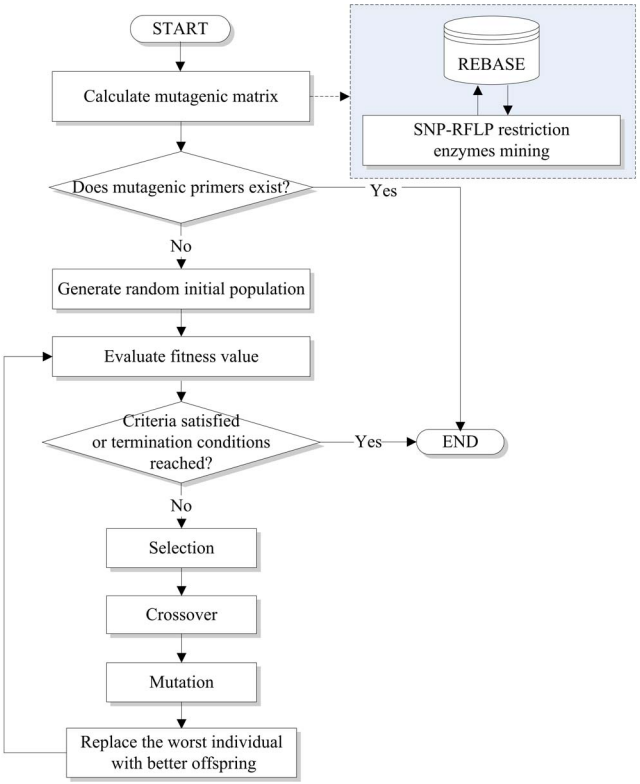

Fig. 3. Flowchart of the GA-based mismatch PCR-RFLP primer design. At first, a random initial population is generated and then all score values of all primer sets in the population are calculated by the score function. A judgment on termination conditions is carried out. If the termination conditions are reached the method stops, else the method proceeds with the following processes. Selection, crossover, and mutation operations are performed and finally the worst primer pairs are replaced by the better primer pairs. The procedure is repeated in the next iteration until the termination conditions are reached.

3. *Creation of a random initial primer set.* Initially, dozens of or hundreds of solutions (P_v) of primer sets, here called a population, are randomly generated without duplicates. F_m is randomly generated between F_{m_min} and F_{m_max}. F_l is randomly generated between the minimum length (16 nt) of the primer and the maximum length (28 nt) of the primer. In order to limit the PCR product length, the proposed method randomly generated P_l between P_{min} and P_{max}. R_l is randomly generated in the same way as F_l.

4. *Evaluation of a primer pair score.* The score value of every primer set is evaluated in turn by an experiential scoring function. In order to let a primer pair satisfy the primer design constraints, the aforementioned primer constraints are used to evaluate the score value. When the primer constraints are all conformed to, the score value is the best value (i.e., zero). Lower scores are the better. The proposed method finds an optimal score from the population by progressing through the iterations. The score value of every primer pair is evaluated in turn by the fitness function (1). In the score function, weights are used to discriminate the significance of each primer constraint function. Five different weight values are used to represent different degrees of importance for these functions; they were 3, 10, 50, 60, and 100, respectively. These weights can be adjusted by researchers based on their own experience or

experiential requirements. Larger weights represent that a function is more important, and vice versa.

$$\begin{aligned} Fitness(P_v) = {} & 3^*(Len_{diff}(P_v) + GC_{proportion}(P_v) \\ & + GC_{clamp}(P_v)) + 10^*(Tm(P_v) + Tm_{diff}(P_v) \\ & + dimer(P_v) + hairpin(P_v)) \\ & + 50^* specificity(P_v) + 60^* product(P_v) \\ & + 100^* RFLP(P_v). \end{aligned} \tag{1}$$

Here, 10 functions are used to estimate the suitability of the designed primer pair; these functions are

$$\begin{aligned} & Len_{diff}(P_v), Tm(P_v), Tm_{diff}(P_v), GC_{proportion}(P_v), \\ & GC_{clamp}(P_v), dimer(P_v), hairpin(P_v), \\ & specificity(P_v), product(P_v), \text{ and } RFLP(P_v) \end{aligned}$$

They are described below:

$Len_{diff}(P_v)$. $Len_{diff}(P_v)$ is used to check whether the length difference of a primer pair exceeds 3 nt (2). Binary values 0 and 1 are applied to represent the two different conditions:

$$Len_{diff}(P_v) = \begin{cases} 0, & \text{if the absolute value of } (F_l - R_l) \le 3, \\ 1, & \text{other conditions.} \end{cases} \tag{2}$$

$Tm(P_v)$. The $Tm(P_v)$ function checks the conditions of melting temperature Tm in a primer pair (3)

$$Tm(Pv) = \begin{cases} 0, & \text{if } 45 \le Tm(P_f) \le 62 \text{ and } 45 \\ & \quad \le Tm(P_r) \le 62 \\ 1, & \text{if } 45 \le Tm(P_f) \le 62 \text{ or } 45 \le Tm(P_r) \le 62 \\ 2, & \text{other conditions.} \end{cases} \tag{3}$$

$Tm_{diff}(P_v)$. The $Tm_{diff}(P_v)$ function checks whether the T_m difference exceeds 5°C (4). The calculation of the melting temperature $Tm(P)$ used here is based on the Wallace's formula (5)

$$Tm_{diff}(P_v) = \begin{cases} 0, & \text{if the absolute value of} \\ & (Tm(P_f) - Tm(P_r)) \le 5, \\ 1, & \text{other conditions,} \end{cases} \tag{4}$$

$$Tm(P) = 4 \times (\#\mathrm{G} + \#\mathrm{C}) + 2 \times (\#\mathrm{A} + \#\mathrm{T}), \tag{5}$$

where #G and #C represent the number of "G" and "C" nucleotides, respectively; #A and #T represent the number of "A" and "T" nucleotides, respectively.

$GC_{proportion}(P_v)$. The GC proportion gives the ratio of the "G" and "C" nucleotides in a primer and can be calculated by $GC\%(P)$ (6). The $GC_{proportion}(P_v)$ function checks whether the $GC\%(P)$ function of the forward and reverse primer lies between 40 and 60 percent (7)

$$GC\%(P) = \frac{G_{number}(P) + C_{number}(P)}{|P|}, \tag{6}$$

where $G_{number}(P)$ represents the number of the nucleotide "G" and $C_{number}(P)$ represents the number of the nucleotide "C."

$$GC_{proportion}(P_v) = \begin{cases} 0, & \text{if } 40\% \le GC\%(P_f) \le 60\% \\ & \quad \text{and } 40\% \le GC\%(P_r) \le 60\% \\ 1, & \text{if } 40\% \le GC\%(P_f) \le 60\% \text{ or} \\ & \quad 40\% \le GC\%(P_r) \le 60\% \\ 2, & \text{other conditions.} \end{cases} \tag{7}$$

$GC_{clamp}(P_v)$. The function $GC_{clamp}(P_v)$ is used to check whether the "G" or "C" nucleotide whether appears at the $3'$ terminal end of a primer; it is defined as follows:

$$GC_{clamp}(P_v) = \begin{cases} 0, & \text{if } P_f \text{ } 3' \text{ end is ``G'' or ``C''} \\ & \quad \text{and } P_r \text{ } 3' \text{ end is ''G'' or ``C''} \\ 1, & \text{if } P_f \text{ } 3' \text{ end is ``G'' or ``C''} \\ & \quad \text{or } P_r \text{ } 3' \text{ end is ``G'' or ``C''} \\ 2, & \text{if neither } P_f \text{ } 3' \text{ end nor } P_r \text{ } 3' \text{ end} \\ & \quad \text{is ``G'' or ''C''.} \end{cases} \tag{8}$$

$dimer(P_v)$. The function $dimer(P_v)$ is applied to check dimer conditions

$$dimer(P_v) = \begin{cases} 0, & \text{if } P_f \text{ and } P_r \text{ do not form a} \\ & \quad \text{primer dimer} \\ 1, & \text{if either } P_f \text{ and } P_r \text{ form a} \\ & \quad \text{cross-dimer or self-dimer} \\ 2, & \text{other conditions.} \end{cases} \tag{9}$$

$hairpin(P_v)$. The function $hairpin(P_v)$ checks for self-annealing in a primer pair and it is defined as

$$hairpin(P_v) = \begin{cases} 0, & \text{if neither } P_f \text{ nor } P_r \text{ forms a hairpin} \\ 1, & \text{if either } P_f \text{ or } P_r \text{ forms a hairpin} \\ 2, & \text{if both } P_f \text{ and } P_r \text{ form a hairpin.} \end{cases} \tag{10}$$

Detailed calculations of dimers/hairpins can be found in Additional file 1, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TCBB.2012.25.

$specificity(P_v)$. The $specificity(P_v)$ function is defined as the number of P_f and P_r that reappear in T_D. It is used to ensure the specificity of the primer pair. A higher value means that a primer pair is less specific

$$specificity(P_v) = \text{the number of } P_f \text{ and } P_r \text{ reappear in } T_D. \tag{11}$$

$product(P_v)$. In order to ensure that the digested allelic fragments can be easily distinguished by electrophoresis, the PCR product size is designed to be between 100 and 300 nt, excluding P_{l1}

$$product(P_v) = \begin{cases} 0, & \text{if product size corresponds} \\ & \quad \text{to the criterion} \\ 1, & \text{otherwise.} \end{cases} \tag{12}$$

$RFLP(P_v)$. The availability $RFLP(P_v)$ of a restriction enzyme is precalculated by a mutagenic matrix. The four values 0, 1, 2, and 3 are returned. The value 0 represents that the designed mutagenic primer has an available restriction enzyme to discriminate the target SNP in both

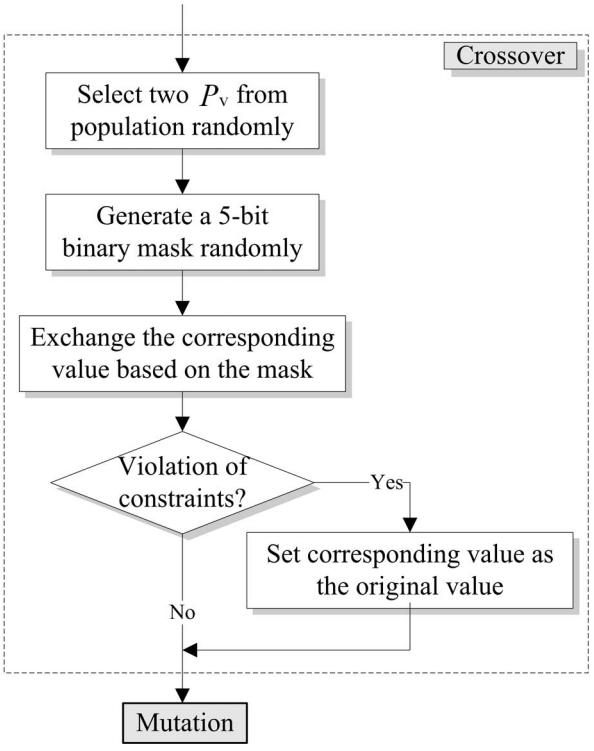

Fig. 4. Crossover flowchart for mismatch PCR-RFLP primer design. Two P_v from the population are randomly selected for crossover. At first, a five bit binary mask is generated which indicates which variables need to be exchanged. All exchanged variables are checked for violation of a constraint. If a constraint is violated, the exchanged variable is restored; else the process proceeds to the next step.

the sense and antisense strand. The value 1 means that the designed mutagenic primer has an available restriction enzyme to discriminate the target SNP in either the sense or antisense strand. The value 2 represents that the designed mutagenic primer has an available restriction enzyme to discriminate the target SNP in neither the sense nor antisense strand. And finally, the value 3 represents that no any mutagenic primer can be designed

$$RFLP(P_v) = mutagenicMetric[F_{m\_\max} - \mathrm{m}][\mathrm{n}]. \quad (13)$$

5. *Judgment of termination criteria.* Termination occurs based on two criteria in this proposed method. If all primer constraints in a primer pair are conformed to, i.e., the score value reaches the best value of zero, the process is terminated. Termination also occurs when a preset number of iterations (generations) is reached. The number of iterations is described in the "Parameter settings" section below.

6. *Operations of selection, crossover, mutation, and replacement.* Selection, crossover, mutation, and replacement operations are the major evolutionary computation processes in the proposed method. The selection operation uses the tournament method. This means that the two best primer pairs in the population (primer sets) are be selected for the crossover and mutation operations. After the two best primer pairs are selected, the crossover operation is then performed. When the set crossover rate is achieved, the selected two primer pairs are processed by the uniform crossover operation and two new primer pairs are generated. The flowchart of the crossover process is shown in Fig. 4, and an example of the crossover operation is shown in Fig. 5. After the crossover operation, the mutation operation is performed. In the mutation operation, one-point mutation is applied in the proposed method. If the set mutation rate is achieved, one offspring is randomly selected after crossover to perform the mutation. The mutation process flowchart is shown in Fig. 6, and an example of the mutation operation is shown in Fig. 7. Finally, the worst two primer pairs are replaced by the new primer pairs via the replacement operation.

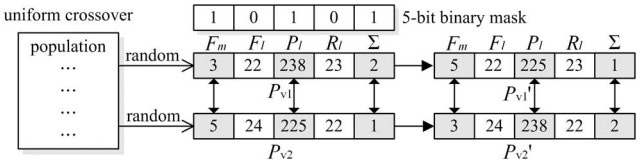

Fig. 5. An example of a crossover operation for mismatch PCR-RFLP primer design. First, two P_v are randomly selected from the population, for example, $P_{v1} = (3, 22, 238, 23, 2)$ and $P_{v2} = (5, 24, 225, 22, 1)$. Then a mask of five binary bits, i.e., 10101, is randomly generated, and based on this mask, the values F_m, R_l, and Σ are exchanged. Finally, the new vectors $P'_{v1} = (5, 22, 225, 23, 1)$ and $P'_{v2} = (3, 24, 238, 22, 2)$ are generated.

3 Results

3.1 Parameter Settings

Four main parameters are set for the proposed algorithm, i.e., the number of iterations (generations), the population size, the crossover rate, and the mutation rate. The

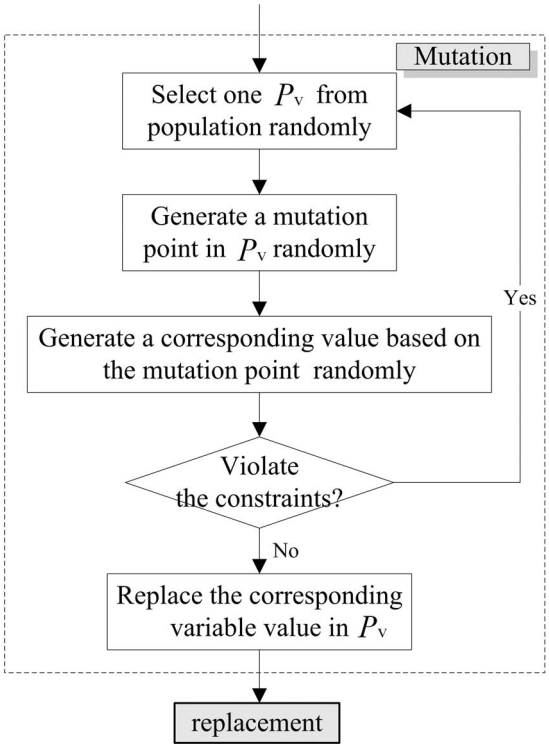

Fig. 6. Mutation flowchart for mismatch PCR-RFLP primer design. When a mutation operation is performed, a mutation point in P_v is randomly selected and a corresponding value generated. Then, the variable is checked for the violation of a constraint. If a constraint is violated, a random point is reselected and the variable is regenerated, otherwise the original value is replaced by the variable and the process continues in the next step.

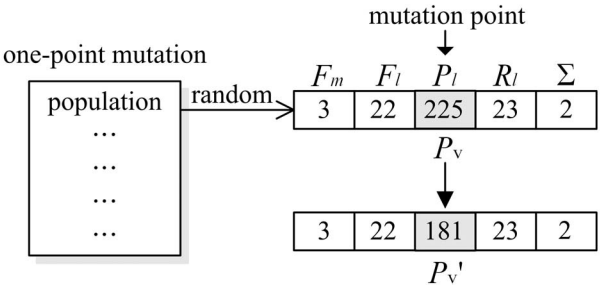

Fig. 7. An example of a mutation operation for mismatch PCR-RFLP primer design. At the beginning, a $P_v = (3, 22, 225, 23, 2)$ is randomly selected from one of the new primer pairs after the crossover operation. Then, one mutation point of the position of F_m, F_l, P_l, R_l or Σ is randomly generated. In our example the position of the mutation point is P_l. Then, a random value between P_{min} and P_{max} is generated, for instance, 181. This value replaces the corresponding variable value in P_v. Finally, the new vector $P'_v = (3, 22, 181, 23, 2)$ is generated.

respective values were 1,000, 50, 0.6, and 0.001; the values are based on DeJong and Spears' parameter settings [35]. Additional parameters are set for different population sizes (100, 200, 300, 400, 500, 600, 700, 800, 900, and 1,000), crossover rates (0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, and 1.0), and mutation rates (0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.09, and 0.10). Furthermore, better parameter values were set based on the simulation results for different population sizes, crossover rates, and mutation rates. These values are 1,000, 900, 0.9 and 0.09, respectively.

3.2 *In Silico* Simulation for GAMPD

Thirty-four SNPs with mismatch PCR-RFLP were detected from the 288 SNPs in the SLC6A4 gene by the proposed method. All designed mismatch PCR-RFLP primer sets are provided at http://bio.kuas.edu.tw/gampd/results.jsp. The statistical numbers for the states of all primers are shown in Tables 1, 3, 4, and 5 (Tables 3, 4, and 5 are shown in Additional file 3, available in the online supplemental material). Sixty-eight primers were designed for the 34 SNPs (34 pairs of primer sets). Therefore, the number of the primer length, GC%, GC clamp, T_m, hairpin and specificity is 68 (a primer contains one GC%, one GC clamp, one T_m, one hairpin and one specificity); the number for the length difference and T_m difference is 34 (a primer pair contains one length difference and one T_m difference); the number for the PCR product length and the dimer is 102 (a primer pair contains three product lengths and three dimers, i.e., one cross-dimer and two self-dimers). The mutagenic primer (P_{l1}) was not considered for the PCR product length due to its limit of 16 and 28 nt, which are not desired PCR product length. The following describes the *in silico* simulation of GAMPD with different parameter settings and the results.

1. *In silico* simulation based on the DeJong and Spears parameter settings

Dejong and Spears' parameter settings are the standard used for most GAs; therefore we used the parameter settings in the present study. Typically, crossover is applied at more than or equal to a rate of 0.6, and the mutation rate is equal to 0.001 [35]. Table 1 (ID1) shows the results of GAMPD based on the DeJong and Spears parameter settings. Six designed primers violated the primer length difference criterion. Most of the primer length differences were between 0 and 3 nt (data not shown). Thirteen primers had a ratio higher than 60 percent; the other ratios were between 40 and 60 percent (data not shown). Approximately 32 percent of the primers (22/68) satisfied the GC clamp restriction. Sixty-five percent of the designed primers $(44/68)$ satisfied the T_m restriction, and 21 percent of the primer pairs $(21/102)$ satisfied the T_m difference criterion. Ninety-six percent of the designed primers $(65/68)$ eliminate the hairpin structure. All designed primers were dimer-free and satisfied the PCR product length and specificity restrictions.

2. *Better simulation of parameter settings in GAMPD.* For the simulations, we used more optimal parameters when executing GAMPD; the results are shown on Table 1 (ID2). The score value was 12.98. Ninety-seven percent of the designed primers $(33/34)$ satisfied the primer length difference restriction. Most of the primer length differences were between 0 and 3 nt (data not shown). Five primers had a GC% of more than 60 percent; the others were in the 40 and 60 percent range (data not shown). Approximately 50 percent of the primers $(34/68)$ satisfied the GC clamp criterion. Approximately 98 percent of the designed primers satisfied the T_m condition $(67/68)$, and 82 percent $(28/34)$ of the designed primers and primer pairs satisfied the T_m difference restriction. All primers satisfied the PCR product length and the dimer criteria. Ninety-five percent

TABLE 1
The Statistical Numbers of Designed Mismatch PCR-RFLP Primers that Satisfy the Common Constraints in 34 SNPs of the SLC6A4 Gene Based on the DeJong and Spears (ID1) and the More Optimal (ID2) Parameter Settings Used by the Proposed Method

	Constraints									
ID	primer length difference	GC%	GC clamp	T_m	T_m difference	product length	dimer	hairpin	specificity	score
1	28/34 (82.4%)	34/68 (50%)	22/68 (32.4%)	44/68 (64.7%)	21/102 (20.6%)	68/68 (100%)	102/102 (100%)	65/68 (95.6%)	68/68 (100%)	14.55
2	33/34 (97.1%)	42/68 (61.8%)	34/68 (50%)	67/68 (98.5%)	28/34 (82.4%)	68/68 (100%)	102/102 (100%)	65/68 (95.6%)	67/68 (98.5%)	12.98

The numerator represents the number of designed mismatch PCR-RFLP primers that satisfy the common constraints and the denominator is the number of all designed mismatch PCR-RFLP primers, respectively. The score value represents the estimated criterion for designing mismatch PCR-RFLP primers The score value is more ideal than for any of the aforementioned results for designing mismatch PCR-RFLP primers. A lower score value is more ideal for designing mismatch PCR-RFLP primers. The score value of the more optimal parameter settings is more ideal than for any of the aforementioned results for designing mismatch PCR-RFLP primers.

TABLE 2
The Characteristics Described in SNP Cutter and GAMPD

Characteristics	SNP Cutter	GAMPD
Mismatch PCR-RFLP SNPs	2	34
No restriction enzymes site were found in sequence with mutagenic location	25	3
Missing in dbSNP	196	0
Unknown	12	0

SNP Cutter only accepts input in specified formats or dbSNP reference IDs. For the SLC6A4 gene, SNP Cutter only designed two mismatch PCR-RFLP SNPs; however, there are 25 SNPs without restriction enzymes found. One hundred and ninety-six SNPs were missing from their databases, and 12 SNPs could not be designed for unknown reasons. In contrast to SNP Cutter, GAMPD detected 37 SNPs without restriction enzymes needed for distinguish, and 34 SNPs were designed for mismatch PCR-RFLP SNPs. Three SNPs without restriction enzymes sites were found in the sequence with mutagenic location.

of the designed primers $(65/68)$ did not show a hairpin structure, and 98 percent of the designed primers $(67/68)$ satisfied the specificity criterion.

3.3 Comparison with SNP Cutter

In order to estimate the effectiveness of GAMPD, we used SNP Cutter to design the mismatch PCR-RFLP primers for the SNPs of the SLC6A4 gene and compared our results to it. We did not consider a comparison with V-MitoSNP and Prim-SNPing (for the reasons given in the DISCUSSION section). The comparison results are shown on Table 2. A detailed comparison for these SNPs is shown in Additional file 2, available in the online supplemental material.

4 Discussion

To date, only a few systems, such as V-MitoSNP, SNP Cutter, and Prim-SNPing, provide a function for mismatch PCR-RFLP primer design to genotype SNPs. However, these systems lack an appropriate algorithm to design feasible mismatch PCR-RFLP primers. Many primer design approaches have been proposed, e.g., dynamic programming [36], genetic algorithm [19], [20], parthenogenetic algorithm MG-PGA [37], greedy algorithm [38], and heuristic algorithm [39]. Nevertheless, most of these methods do not focus on the design of mismatch PCR-RFLP primers nor provide the available restriction enzymes for SNP genotyping. In this paper, a useful mismatch PCR-RFLP primer design method is proposed to facilitate PCR-RFLP. A total of 288 SNPs of the SLC6A4 gene were used and 34 SNPs were successfully designed with GAMPD.

1. *GAMPD for mismatch PCR-RFLP primer design.* In mismatch PCR-RFLP for SNP genotyping, information about the mutagenic primer and restriction enzymes is most important. Since the original sequence does not contain any available enzymes to distinguish a target SNP, the mutagenic primer is introduced to find available restriction enzymes for the PCR-RFLP experiments. If no enzyme is available for the mutagenic primer to distinguish the target SNP, the PCR-RFLP experiment is meaningless. In our method, the updated core of the SNP-RFLPing [26], [27], [28] program is used to mine for available restriction enzymes. A mutagenic matrix is applied to record the mutagenic states effectively for the further operations. GAMPD usually converges all solutions into one optimal solution in a short time by using crossover and mutation. When designing the mismatch PCR-RFLP primer pair, a minimum sequence length is required. Hence, the SNP-Flankplus [33] program is used to obtain flanking sequences of 500 nt in length for a target SNP. The typical primer design constraints for primer length, length difference, T_m, T_m difference, GC proportion, GC clamp, PCR product length are used by GAMPD. Moreover, secondary structures, such as dimer formation (including cross-dimer and self-dimer), hairpin structure, and the specificity in a template sequence were also applied to obtain feasible mismatch PCR-RFLP primers.

2. *Influence of the parameter settings.* The GA is capable of finding PCR-RFLP primers which correspond to the primer constraints. The *in silico* simulation of the proposed GAMPD showed that it reliably fits the constraints to the primers (Table 1 and Additional file 3, available in the online supplemental material). The Dejong and Spears parameter settings are the standard used for most GAs and for this reason were used in the present study. Typically, crossover is applied at more than or equal to the rate of 0.6, and the mutation rate is usually equal to 0.001 [35]. However, the population size of 50 used by Dejong and Spears is too small to provide the necessary sampling accuracy for the design of mismatch PCR-RFLP primer sets. Consequently, we increased the population size from 50 to 100, 200, 300, 400, 500, 600, 700, 800, 900, and 1,000 to ensure more accurate sampling. In addition, in order to observe the influences of the crossover rates and the mutation rates, we changed these rates to design the mismatch PCR-RFLP primers. The standard deviation (SD) of the results shows the influence of the constraints in different settings of the parameter (see the Additional file 3, available in the online supplemental material).

3. *Availability of GAMPD.* GAMPD can be accessed at http://bio.kuas.edu.tw/gampd/. It designs appropriate mismatch PCR-RFLP primers for genotyping a defined SNP based on default parameter settings (number of iterations, population size, crossover rate, and mutation rate are 1,000, 900, 0.9, and 0.09, respectively); these settings achieved the best simulation for the SLC6A4 gene in Table 1 (ID2). However, parameter settings or the primer design

conditions can be individually changed by users based on their requirements. When the input sequence contains multiple SNPs, the first SNP will be taken as the target SNP. The input SNP format can be [dNTP1/dNTP2] or IUPAC code. Eventually, a set of available restriction enzymes and a feasible mismatch PCR-RFLP primer pair are reported in a text file that records all relevant information of the mismatch PCR-RFLP primer pair.

5 Conclusions

V-MitoSNP and Prim-SNPing are two web-based programs containing the function of PCR-RFLP primer design. However, V-MitoSNP only focuses on mitochondrial SNPs and does not provide sequence or SNP ID input functions, thus limiting its further application for designing PCR-RFLP primers. Although Prim-SNPing provides a user-friendly interface, it uses a window-sliding strategy which limits its search efficiency and results. Furthermore, designed primers are strict (all primer constraints must be conformed to) and therefore it usually does not get approximate outputs. SNP Cutter uses the popular primer design program Primer3 to provide suitable primers and identifies available restriction enzymes. However, it is limited by the specified input formats. Many SNP IDs are not found, however, it does not provide information regarding the latest restriction enzymes. It thus missed of many feasible solutions (Table 2). GAMPD uses the updated core of the SNP-RFLPing program to mine entire REBASE and applies a GA, evolutionary computation to design mismatch PCR-RFLP primers. It avoids insufficient restriction enzymes and eliminates absolute primer constraints and therefore feasible mismatch PCR-RFLP primers can be obtained. In conclusion, the proposed GAMPD is a reliable method for designing feasible mismatch PCR-RFLP primers since it conforms to most of the primer constraints and provides entire restriction enzymes information. The test flexibility of GAMPD has been demonstrated for many polymorphisms, i.e., 34 SNPs in SLC6A4 gene.

Acknowledgments

This work is partly supported by the National Science Council in Taiwan under grant. NSC96-2221-E-214-050-MY3, NSC98-2221-E-151-040-, NSC98-2622-E-151-001-CC2 and NSC98-2622-E-151-024-CC3.

References

[1] R. Zhang, Z. Zhu, H. Zhu, T. Nguyen, F. Yao, K. Xia, D. Liang, and C. Liu, "SNP Cutter: A Comprehensive tool for SNP PCR-RFLP Assay Design," *Nucleic Acids Research,* vol. 33, pp. W489-W492, July 2005.

[2] M. Ota, H. Fukushima, J.K. Kulski, and H. Inoko, "Single Nucleotide Polymorphism Detection by Polymerase Chain Reaction-Restriction Fragment Length Polymorphism," *Nature Protocols,* vol. 2, pp. 2857-2864, 2007.

[3] J.G. Hacia, J.B. Fan, O. Ryder, L. Jin, K. Edgemon, G. Ghandour, R.A. Mayer, B. Sun, L. Hsie, C.M. Robbins, L.C. Brody, D. Wang, E.S. Lander, R. Lipshutz, S.P. Fodor, and F.S. Collins, "Determination of Ancestral alleles for Human Single-Nucleotide Polymorphisms Using High-Density Oligonucleotide Arrays," *Nature Genetics,* vol. 22, pp. 164-167, June 1999.

[4] V. Mooser, D.M. Waterworth, T. Isenhour, and L. Middleton, "Cardiovascular Pharmacogenetics in the SNP era," *J. Thrombosis and Haemostasis,* vol. 1, pp. 1398-1402, July 2003.

[5] H.C. Erichsen and S.J. Chanock, "SNPs in Cancer Research and Treatment," *British J. Cancer,* vol. 90, pp. 747-751, 2004.

[6] L.J. Engle, C.L. Simpson, and J.E. Landers, "Using High-Throughput SNP Technologies to Study Cancer," *Oncogene,* vol. 25, pp. 1594-1601, 2006.

[7] C.R. Cantor, "The Use of Genetic SNPs as New Diagnostic Markers in Preventive Medicine," *Annals of the New York Academy of Sciences,* vol. 1055, pp. 48-57, 2005.

[8] T. Bernig and S.J. Chanock, "Challenges of SNP Genotyping and Genetic Variation: Its Future Role in Diagnosis and Treatment of Cancer," *Expert Review of Molecular Diagnostics,* vol. 6, pp. 319-331, 2006.

[9] W. Sadee, "Pharmacogenomics: Harbinger for the Era of Personalized Medicine?," *Molecular Interventions,* vol. 5, pp. 140-143, 2005.

[10] A.M. Divne and M. Allen, "A DNA Microarray System for Forensic SNP Analysis," *Forensic Science Int'l,* vol. 154, pp. 111-121, 2005.

[11] R.J. Roberts, T. Vincze, J. Posfai, and D. Macelis, "REBASE-Enzymes and Genes for DNA Restriction and Modification," *Nucleic Acids Research,* vol. 35, pp. D269-D270, Jan. 2007.

[12] L.Y. Chuang, C.H. Yang, Y.H. Cheng, D.L. Gu, P.L. Chang, K.H. Tsui, and H.W. Chang, "V-MitoSNP: Visualization of Human Mitochondrial SNPs," *BMC Bioinformatics,* vol. 7, article 379, 2006.

[13] S. Rozen and H. Skaletsky, "Primer3 on the WWW for General Users and for Biologist Programmers," *Methods in Molecular Biology,* vol. 132, pp. 365-386, 2000.

[14] T. Koressaar and M. Remm, "Enhancements and Modifications of Primer Design Program Primer3," *Bioinformatics,* vol. 23, pp. 1289-91, May 2007.

[15] F.M. You, N. Huo, Y.Q. Gu, M.C. Luo, Y. Ma, D. Hane, G.R. Lazo, J. Dvorak, and O.D. Anderson, "BatchPrimer3: A High throughput Web Application for PCR and Sequencing Primer Design," *BMC Bioinformatics,* vol. 9, article 253, 2008.

[16] H.W. Chang, L.Y. Chuang, Y.H. Cheng, Y.C. Hung, C.H. Wen, D.L. Gu, and C.H. Yang, "Prim-SNPing: A Primer Designer for Cost-Effective SNP Genotyping," *Biotechniques,* vol. 46, pp. 421-31, May 2009.

[17] M.J. McPherson, G.R. Taylor, and P. Quirke, *PCR, a Practical Approach.* Oxford Univ. Press, 1991.

[18] J. Sambrook and D.W. Russell, *Molecular Cloning: A Laboratory Manual.* Cold Spring Harbor Laboratory Press, 2001.

[19] S.H. Chen, C.Y. Lin, C.S. Cho, C.Z. Lo, and C.A. Hsiung, "Primer Design Assistant (PDA): A Web-Based Primer Design Tool," *Nucleic Acids Research,* vol. 31, pp. 3751-3754, July 2003.

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

[21] C.H. Yang, Y.H. Cheng, L.Y. Chuang, and H.W. Chang, "Specific PCR Product Primer Design Using Memetic Algorithm," *Biotechnol Prog,* vol. 25, pp. 745-53, May/June 2009.

[22] A. Haliassos, J.C. Chomel, S. Grandjouan, J. Kruh, J.C. Kaplan, and A. Kitzis, "Detection of Minority Point Mutations by Modified PCR Technique: A New Approach for a Sensitive Diagnosis of Tumor-Progression Markers," *Nucleic Acids Research,* vol. 17, pp. 8093-8099, Oct. 1989.

[23] A. Haliassos, J.C. Chomel, L. Tesson, M. Baudis, J. Kruh, J.C. Kaplan, and A. Kitzis, "Modification of Enzymatically Amplified DNA for the Detection of Point Mutations," *Nucleic Acids Research,* vol. 17, p. 3606, May 1989.

[24] K.D. Jong, "Learning with Genetic Algorithms: An Overview," *Machine Learning,* vol. 3, pp. 121-138, 1988.

[25] D.E. Goldgerg, *Genetic Algorithms in Search, Optimization, and Machine Learning.* Addison-Wesley 1989.

[26] H.W. Chang, C.H. Yang, P.L. Chang, Y.H. Cheng, and L.Y. Chuang, "SNP-RFLPing: Restriction Enzyme mining for SNPs in Genomes," *BMC Genomics,* vol. 7, article 30, 2006.

[27] L.Y. Chuang, C.H. Yang, K.H. Tsui, Y.H. Cheng, P.L. Chang, C.H. Wen, and H.W. Chang, "Restriction Enzyme Mining for SNPs in Genomes," *Anticancer Research,* vol. 28, pp. 2001-2007, July/Aug. 2008.

[28] H.W. Chang, Y.H. Cheng, L.Y. Chuang, and C.H. Yang, "SNP-RFLPing 2: An Updated and Integrated PCR-RFLP Tool for SNP Genotyping," *BMC Bioinformatics,* vol. 11, article 173, Apr. 2010.

[29] J. Holland, *Adaptation in Nature and Artificial Systems.* MIT Press, 1992.

[30] T. Sakurai, J. Reichert, E.J. Hoffman, G. Cai, H.B. Jones, M. Faham, and J.D. Buxbaum, "A Large-Scale Screen for Coding Variants in SERT/SLC6A4 in Autism Spectrum Disorders," *Autism Research,* vol. 1, pp. 251-257, Aug. 2008.
[31] T.E. Goldberg, R. Kotov, A.T. Lee, P.K. Gregersen, T. Lencz, E. Bromet, and A.K. Malhotra, "The Serotonin Transporter Gene and Disease Modification in Psychosis: Evidence for Systematic Differences in Allelic Directionality at the 5-HTTLPR Locus," *Schizophr Research,* vol. 111, pp. 103-108, June 2009.
[32] L. Mandelli, M. Mazza, G. Martinotti, M. Di Nicola, D. Tavian, E. Colombo, S. Missaglia, D. De Ronchi, R. Colombo, and L. Janiri, "Harm Avoidance Moderates the Influence of Serotonin Transporter Gene Variants on Treatment Outcome in Bipolar Patients," *J. Affective Disorders,* vol. 119, pp. 205-209, 2009.
[33] C.H. Yang, Y.H. Cheng, L.Y. Chuang, and H.W. Chang, "SNP-Flankplus: SNP ID-Centric Retrieval for SNP Flanking Sequences," *Bioinformation,* vol. 3, pp. 147-149, 2008.
[34] S.T. Sherry, M.H. Ward, M. Kholodov, J. Baker, L. Phan, E.M. Smigielski, and K. Sirotkin, "dbSNP: The NCBI Database of Genetic Variation," *Nucleic Acids Research,* vol. 29, pp. 308-311, Jan. 2001.
[35] K.A. De Jong and W.M. Spears, "An Analysis of the Interacting Roles of Population Size and Crossover in Genetic Algorithms," *Proc. First Workshop Parallel Problem Solving from Nature,* vol. 496, pp. 38-47, 1990.
[36] T. Kampke, M. Kieninger, and M. Mecklenburg, "Efficient Primer Design Algorithms," *Bioinformatics,* vol. 17, pp. 214-225, Mar. 2001.
[37] J. Wu, J. Wang, and J. Chen, "A Practical Algorithm for Multiplex PCR Primer Set Selection," *Int'l J. Bioinformatics Research and Applications,* vol. 5, pp. 38-49, 2009.
[38] J. Wang, K.B. Li, and W.K. Sung, "G-PRIMER: Greedy Algorithm for Selecting Minimal Primer Set," *Bioinformatics,* vol. 20, pp. 2473-2475, Oct. 2004.
[39] Y.F. Chen, R.C. Chen, Y.K. Chan, R.H. Pan, Y.C. Hseu, and E. Lin, "Design of Multiplex PCR Primers Using Heuristic Algorithm for Sequential Deletion Applications," *Computational Biology and Chemistry,* vol. 33, pp. 181-8, Apr. 2009.

Cheng-Hong Yang received the MS and PhD degrees in computer engineering from North Dakota State University in 1990 and 1992, respectively. He is a professor in the Department of Electronic Engineering at National Kaohsiung University of Applied Sciences and serves as a president of the Toko University in Chiayi, Taiwan. His main areas of research are evolutionary computation, bioinformatics, and assistive tool implementation.

Yu-Huei Cheng received the MS and PhD degrees from the Department of Electronic Engineering, National Kaohsiung University of Applied Sciences, Taiwan, in 2006 and 2010, respectively. He has rich experiences in computer programming, database design and management, and systems programming and design. His main areas of research are bioinformatics and computational biology.

Cheng-Huei Yang received the BS degree and the MS degree from National Taipei Institute of Technology and Northeastern University, in 1978 and 1987, respectively. He received the PhD degree in electrical engineering from National Cheng Kung University in 2001. Currently, he is an associate professor in the Department of Telecommunication and Computer Engineering, National Kaohsiung Institute of Marine Technology. His current research interests include network communication, electronic instrument systems, and image processing.

Li-Yeh Chuang received the MS degree from the Department of Chemistry at University of North Carolina in 1989 and the PhD degree from Department of Biochemistry at North Dakota State University in 1994. She is a professor and director of the Department of Chemical Engineering & Institute of Biotechnology and Chemical Engineering at I-Shou University, Kaohsiung, Taiwan. Her main areas of research are bioinformatics, biochemistry, and genetic engineering.