Confronting two-pair primer design for enzyme-free SNP genotyping based on a genetic algorithm

Cheng-Hong Yang1,2, Yu-Huei Cheng1, Li-Yeh Chuang3*, Hsueh-Wei Chang4,5,6,7*

Abstract

Background: Polymerase chain reaction with confronting two-pair primers (PCR-CTPP) method produces allele-specific DNA bands of different lengths by adding four designed primers and it achieves the single nucleotide polymorphism (SNP) genotyping by electrophoresis without further steps. It is a time- and cost-effective SNP genotyping method that has the advantage of simplicity. However, computation of feasible CTPP primers is still challenging.

Results: In this study, we propose a GA (genetic algorithm)-based method to design a feasible CTPP primer set to perform a reliable PCR experiment. The SLC6A4 gene was tested with 288 SNPs for dry dock experiments which indicated that the proposed algorithm provides CTPP primers satisfied most primer constraints. One SNP rs12449783 in the SLC6A4 gene was taken as an example for the genotyping experiments using electrophoresis which validated the GA-based design method as providing reliable CTPP primer sets for SNP genotyping.

Conclusions: The GA-based CTPP primer design method provides all forms of estimation for the common primer constraints of PCR-CTPP. The GA-CTPP program is implemented in JAVA and a user-friendly input interface is freely available at http://bio.kuas.edu.tw/ga-ctpp/.

Background

Genotyping is a common technique used in association studies of diseases and cancers. Although many high-throughput platforms of single nucleotide polymorphism (SNP) genotyping, such as SNP array [1] and real-time PCR using TaqMan probes [2], have been introduced, most laboratories still validate SNP or novel mutation by PCR-restriction fragment length polymorphism (RFLP) genotyping [3-6] because this method is inexpensive for small-scale genotyping. One shortcoming of PCR-RFLP is its long digestion time (usually in 2-3 hours) for restriction enzymes [7,8].

Recently, a restriction enzyme-free SNP genotyping technique called “PCR with confronting two-pair primers (PCR-CTTP)” was developed [9-12]. It has been applied successfully to at least 30 different SNP genotypings. For example, interleukin-1B (IL-1B) C-31T, interleukin-2 (IL-2) -330G, beta2-adrenergic receptor (beta2-AR) Gln27Glu, and aldehyde dehydrogenase 2 (ALDH2) were genotyped by PCR-CTPP for association studies with smoking behavior [13], pylori-induced gastric atrophy [14], severe coronary artery disease [15], and esophageal cancer risk [16], respectively. There is no doubt that the PCR-CTPP method is suitable and reliable for most cases of SNPs. This method considerably lowers the need to consume restriction enzymes. However, the criteria for the PCR-CTPP primers are only tolerant of a small difference in melting temperature ($T_{m}$-diff) between the four primers in the PCR-CTPP method [10]. Moreover, typical primer design constraints also need to be considered, such as primer length, primer length difference, PCR product length, GC proportion, melting temperature ($T_{m}$), GC clamp, dimer (including cross-dimers and self-dimers), hairpin structure, and specificity. Hence, the computational requirements needed to improve the primer design with PCR-CTTP are rather high.
To design CTPP primers with many corresponding constraints, we introduce a genetic algorithm (GA) [17,18] to improve the design of CTPP primer sets. GA is a stochastic search algorithm modeled on the process of natural selection underlying biological evolution [19]. It constitutes a randomized search and an optimization technique that derives its working principle from natural genetics. Since GA computation is similar in nature to the evolution of the species, it best fits DNA behavior associated with SNP interaction [20] and general primer design [21]. The evolutionary computations involved, such as selection, crossover and mutation, are effective in achieving optimal solutions for many CTPP primer sets. After each run, chromosomes in a GA share information with each other and the superior solutions based on a fitness rule are refined from generation to generation. Therefore, CTPP primers obeying the typical primer design constraints described above can be mined.

**Methods**

**Problem formulation**

The CTPP primer design problem can be described as follows. Let \( T_D \) be a template DNA sequence, which is composed of nucleotide codes with an identified SNP. \( T_D \) is defined by:

\[
T_D = \{ B_i | \ i \ is \ the \ index \ of \ DNA \ sequence, \ 1 \leq i \leq t, \ \exists \ B_i \in IUPAC \ code \ of \ SNP \} \tag{1}
\]

where \( B_i \) is the regular nucleotide code (A, T, C, or G) mixed with a single IUPAC code of SNP (M, R, W, S, Y, K, V, H, D, B or N) (is the existence and uniqueness). For the target SNP, we focused only on true SNPs (K, V, H, D, B or N) (is the existence and uniqueness). The regular primer design constraints are used as values in GA and is used to perform evolutionary computations as described in the following sections.

**Definition of the fitness function**

The regular primer design constraints are used as values to design a fitness function to minimize the fitness value. The fitness function is defined as follows:

\[
Fitness(P_v) = 3 * (Len_{diff}(P_v)) + GC_{\ proportion}(P_v) + GC_{\ clamp}(P_v)) + 10 * (dimer(P_v)) + hairpin(P_v) + specificity(P_v) + 50 * (Tm(P_v) + Tm_{diff}(P_v)) + 100 * v_g Tm_{diff}(P_v) + 60 * \text{PCRLen}(P_v) \tag{7}
\]

The weights (3, 10, 50, 60 and 100) of the fitness function are applied to estimate the importance of the primer constraints. These weights are set according to the experiential conditions for PCR-CTPP. They also accept adjustment based on the user experimental requirements.

**Primer length**

The feasible primer length for a PCR experiment is set between 16 and 28 bp. For longer primers, the \( T_m \) is
increased whereas the $T_m$ of relatively short primers is decreased. Accordingly, primers which are neither too long nor too short are suitable. We have limited the random values of $F_{11}, R_{11}, F_{12}$ and $R_{12}$ in an appropriate range; therefore, the primer length estimation is not considered to be joined to the fitness function.

A length difference ($L_{\text{diff}}$) of less than or equal to 3 bp between the $F_{11}/R_{11}, F_{12}/R_{12}$ and $F_{13}/R_{13}$ primer sets is considered optimal. The primer length difference function is defined as follows:

$$L_{\text{diff}}(P) = \begin{cases} \text{defect\_value} = 3 \\
\text{if } \text{ABS}(P_{11} - R_{11}) \leq 3, \\
\text{then } \text{defect\_value} - 1 \end{cases}$$

$$L_{\text{diff}}(P) = \begin{cases} \text{defect\_value} = 3 \\
\text{if } \text{ABS}(P_{12} - R_{12}) \leq 3, \\
\text{then } \text{defect\_value} - 1 \end{cases}$$

$$L_{\text{diff}}(P) = \begin{cases} \text{defect\_value} = 3 \\
\text{if } \text{ABS}(P_{11} - R_{12}) \leq 3, \\
\text{then } \text{defect\_value} - 1 \end{cases}$$

$$\text{return defect\_value}$$

where $L_{\text{diff}}(P)$ has a maximal fitness value of 3; the fitness value is decreased when the length difference between a primer pair is less than or equal to 3 bp. ABS represents the absolute value.

**GC content and GC clamp**

The function $GC\%(P)$ is proposed to represent the ratio of G and C nucleotides appearing in a primer:

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

where $P$ represents a primer and $|P|$ represents the length of primer $P$; $G_{\text{number}}(P)$ and $C_{\text{number}}(P)$ represent the numbers of the nucleotides G and C, respectively.

In general primer design, the typical GC proportion constraint is set between 40% and 60%. However, the designed CTPP primers contain the target SNP to limit the range of the GC proportion. To relax this constraint, the constraint of GC proportion in a primer is adjusted to between 20% and 80%. Function $GC_{\text{proportion}}(P)$ is proposed with a maximal fitness value of 4 to lead the $GC\%(P)$ of CTPP primers corresponding to this constraint:

$$GC_{\text{proportion}}(P) = \begin{cases} \text{defect\_value} = 4 \\
\text{if } 20 \leq GC\%(P_{11}) \leq 80, \\
\text{then } \text{defect\_value} - 1 \end{cases}$$

$$GC_{\text{proportion}}(P) = \begin{cases} \text{defect\_value} = 4 \\
\text{if } 20 \leq GC\%(P_{12}) \leq 80, \\
\text{then } \text{defect\_value} - 1 \end{cases}$$

$$GC_{\text{proportion}}(P) = \begin{cases} \text{defect\_value} = 4 \\
\text{if } 20 \leq GC\%(P_{12}) \leq 80, \\
\text{then } \text{defect\_value} - 1 \end{cases}$$

To meet the presence of G or C nucleotides at the 3' terminal of a primer to ensure a tight localized hybridization bond, the function $GC_{\text{clamp}}(P)$ is proposed with the maximal fitness value of 4 as follows:

$$GC_{\text{clamp}}(P) = \begin{cases} \text{defect\_value} = 4 \\
\text{if } 3'\text{ end of } P_{11} \text{ is 'G' or 'C'}, \\
\text{then } \text{defect\_value} - 1 \end{cases}$$

$$GC_{\text{clamp}}(P) = \begin{cases} \text{defect\_value} = 4 \\
\text{if } 3'\text{ end of } P_{12} \text{ is 'G' or 'C'}, \\
\text{then } \text{defect\_value} - 1 \end{cases}$$

$$GC_{\text{clamp}}(P) = \begin{cases} \text{defect\_value} = 4 \\
\text{if } 3'\text{ end of } P_{12} \text{ is 'G' or 'C'}, \\
\text{then } \text{defect\_value} - 1 \end{cases}$$

$$\text{return defect\_value}$$
the PCR-CTPP is less developed for its computational tool providing PCR-CTPP primer design. A novel strategy for PCR-CTPP primer design has been introduced in this paper and the freely available web server implemented with this method was also constructed. With experimental validation, our proposed GA-based method is a useful algorithm to design feasible CTPP primers and it conforms to most of the PCR-CTPP constraints.

Availability and requirements

Project name: GA-CTPP: Confronting two-pair primer design using genetic algorithm.


Operating system(s): Operating systems free with web browser.

Programming language: Java.

Other requirements: Java 1.5.

License: none for academic users. For any restrictions regarding the use by non-academics please contact the corresponding author.

Additional material

Additional file 1: ‘The differences between our previous publication in BIBE 2009 conference [34] and this study’

Additional file 2: ‘The performances for primer design using our proposed GA-CTPP algorithm between different population sizes of Dejong and Spears’s parameter settings’

Acknowledgements


Authors’ contributions

C-HY coordinated and oversaw this study. Y-HC participated in the design of the algorithm, and wrote the program and the manuscript. L-YC provided the biochemistry background and introduced the bioinformatics needed for primer design. H-WC performed and verified the PCR experiment, and modified the manuscript. All authors read and approved the final manuscript.

Received: 9 January 2010 Accepted: 13 October 2010 Published: 13 October 2010