Confronting Two-Pair Primer Design using Memetic Algorithm

Cheng-Hong Yang[1,2], Yu-Huei Cheng[1], and Li-Yeh Chuang[3]
[1]Department of Electronic Engineering, National Kaohsiung University of Applied Sciences, Kaohsiung, Taiwan.
[2]Department of Network Systems, Toko University, Chiayi, Taiwan.
[3]Department of Chemical Engineering & Institute of Biotechnology and Chemical Engineering, I-Shou University, Kaohsiung, Taiwan.
C-HY: chyang@cc.kuas.edu.tw; Y-HC: yuhuei.cheng@gmail.com; L-YC: chuang@isu.edu.tw.

Abstract

Polymerase chain reaction with confronting two-pair primers (PCR-CTPP) is a simple, time- and cost-effective SNP genotyping method. In this study, we propose a MA (memetic algorithm)-based method to design a feasible CTPP primer set. Overall, 288 SNPs in the SLC6A4 gene were tested in silico by the proposed method. The results indicate that the proposed method provides feasible CTPP primers that conform to the commonly used constraints. This method could assist biologists and other researchers in obtaining a feasible CTTP primer set.

1 Introduction

Genotyping is a common technique used in association studies of diseases and cancers. 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, but most laboratories still validate SNPs or novel mutations by PCR-restriction fragment length polymorphism (RFLP) genotyping [3-5] because this method is inexpensive and ideally suited for small-scale genotyping. However, one shortcoming of PCR-RFLP is the long digestion time (usually 2-3 hours) of the restriction enzymes [6].

PCR with confronting two-pair primers (PCR-CTPP) is a restriction enzyme-free SNP genotyping technique developed recently [7]. Many SNPs have been genotyped successfully by it. Examples are the interleukin-1B (IL-1B) C-31T, interleukin-2 (IL-2) -330G, beta2-adrenergic receptor (beta2-AR) Gln27Glu, aldehyde dehydrogenase 2 (ALDH2), all of which were successfully genotyped by PCR-CTPP in association studies of smoking behavior [8], pylori-induced MAstric atrophy [9], severe coronary artery disease [10], and esophageal cancer risk [11], respectively. The method considerably lowers the need to consume restriction enzymes. However, the method is only tolerant of a small difference in melting temperature ($T_{m\text{-}diff}$) between the four primers [12]. Moreover, typical primer design constraints need also be considered, such as primer length, difference of primer pair length, PCR product length, GC proportion, melting temperature (T_m), GC clamp, the existence of dimers (including cross-dimers and self-dimers), hairpin structure, and the specificity.

To design CTPP primers corresponding to the many constraints, we introduced a genetic algorithm (GA) to design CTPP primer sets in the past [13]. However, the T_m difference is not desirable in most cases, and we thus propose a memetic algorithm (MA) [14] to improve the design of CTPP primer sets. MAs were inspired by Dawkins' notion of a meme [15]. A MA assures that all chromosomes and offsprings gain some experiences. Since the computation of MAs is similar the one of GAs, the evolutionary steps involved, such as selection, crossover and mutation, are effective in achieving optimal solutions for many CTPP primer sets. After each run, chromosomes in MA share information with each other and the superior solutions based on a fitness rule are refined from generation to generation. The usage of MAs constitutes a more correct and accurate method to solve CTPP primer design problems, and saves researchers a considerable amount of time and materials when performing PCR-CTPP experiments.

2 Methods

Problem formulation

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

$$T_D = \{B_i \mid i \text{ is the index of DNA sequence}, \; 1 \le i \le \iota, \exists! B_i \in \text{IUPAC code of SNP}\} \quad (1)$$

where B_i is the regular nucleotide code (A, T, C, or

G) mixed with a single IUPAC code of the SNP (M, R, W, S, Y, K, V, H, D, B or N) ($\exists!$ is the existence and uniqueness).

CTPP primer design requires two pairs of short sequences which are constrained to T_D based on a defined SNP site as illustrated in Fig. 1. The forward primer 1 (P_{f1}) is a short sense sequence in the upstream (5' end) of a defined SNP site, the reverse primer 1 (P_{r1}) is a short antisense sequence which contains a nucleotide (the minor allele of the defined SNP site) located at its 3' end, the forward primer 2 (P_{f2}) is a short sense sequence which contains a nucleotide (the major allele of the defined SNP site) located at its 3' end, and the reverse primer 2 (P_{r2}) is the antisense sequence in the upstream of a defined SNP site.

The SNP site is defined at the end positions of P_{f2} and P_{r1}, which are indicated by the symbols F_{e2} and R_{e1}, respectively. As illustrated in Fig. 1, a vector (v) with F_{l1}, P_{l1}, R_{l1}, F_{l2}, P_{l2} and R_{l2} can be created to design the CTPP primer sets. This vector is defined as follows:

$$P_v = (F_{l1}, P_{l1}, R_{l1}, F_{l2}, P_{l2}, R_{l2}) \tag{2}$$

F_{l1}, P_{l1}, R_{l1}, F_{l2}, P_{l2} and R_{l2} represent the number of nucleotides of the forward primer 1, product length between P_{f1} and P_{r1}, reverse primer 1, forward primer 2, product length between P_{f2} and P_{r2} and reverse primer 2, respectively. Consequently, the forward and the reverse primers can be acquired from P_v, which is the prototype of a chromosome in MA, and is used to perform evolutionary computations as described in the following sections.

Fitness function

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

$$\begin{aligned} Fitness(P_v) = {} & 3*(Len_{diff}(P_v)+GC_{proportion}(P_v)+GC_{clamp}(P_v)) \\ & +10*(dimer(P_v)+ hairpin(P_v) + specificity(P_v)) \\ & +50*(Tm(P_v)+Tm_{diff}(P_v))+100*Avg\_Tm_{diff}(P_v) \\ & +60*PCRlen_{ratio}(P_v) \end{aligned} \tag{3}$$

The weights of the fitness function are applied to estimate the importance of the primer constraints. These weights are set based on the experiential conditions for PCR-CTPP. They also accept adjustments based on the experimental requirements.

The feasible primer length for a PCR experiment is set to between 16 and 28 bp. For longer primers, the T_m is increased, whereas the T_m of relatively short primers is decreased. Accordingly, neither primer that is too long nor too short is suitable. Since the random values of F_{l1}, R_{l1}, F_{l2} and R_{l2} are already limited by the constraints, the primer length estimation is not considered by the fitness function. A length difference (Len_{diff}) of less than or equal to 3 bp between the F_{l1}/R_{l1}, F_{l2}/R_{l2}, and F_{l2}/R_{l1} primer sets is considered optimal.

The function $GC\%(P)$ is proposed to represent the G and C nucleotide ratio appearing in a primer. In general primer design, the typical GC proportion constraint is set to between 40% and 60%. However, the designed CTPP primers contain the target SNP have limited the range of the GC proportion. To relax this constraint, the GC proportion in a primer is adjusted to between 20% and 80%. The function $GC_{proportion}(P_v)$ is proposed with a maximal fitness value of 4 so that the $GC\%(P)$ of CTPP primers corresponds to this constraint.

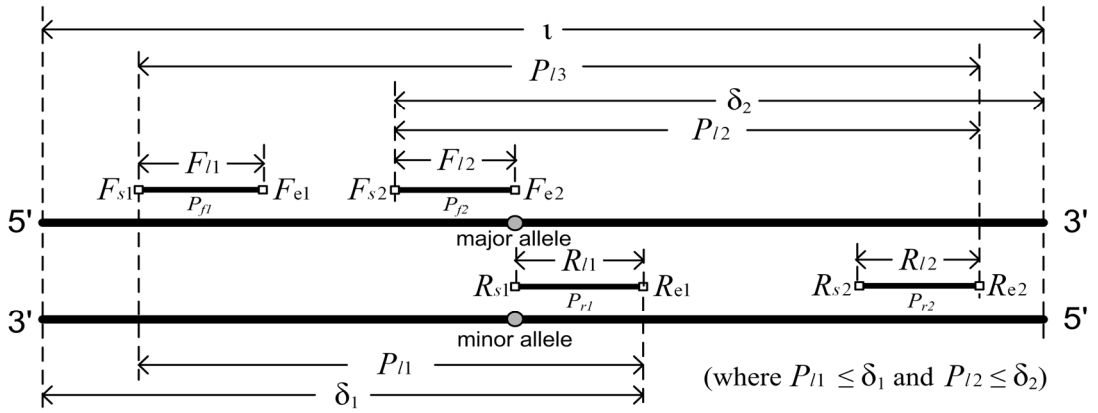

Figure 1. Parameters for the CTPP primer set.

To account for the presence of G or C at the 3' terminal of a primer, the function $GC_{clamp}(P_v)$ is proposed. The melting temperature (T_m) for each CTPP primer must be considered carefully in a PCR experiment. The $Tm(P_v)$ function confines the T_m of a CTPP primer set to range between 45°C and 62°C. Similar T_m values of a primer pair is important for the experiment. The $Tm_{diff}(P_v)$ function ensures that the T_m difference is within a certain limit. In order to balance the T_m values among CTPP primers, the $Avg\_Tm_{diff}(P_v)$ function is proposed to calculate the average T_m difference. The $dimer(P_v)$ and $hairpin(P_v)$ functions check for the existence of dimmers or hairpin structures in the primer pair. The $specificity(P_v)$ function is proposed to check for repetition of each CTPP primer in the template DNA sequence to ensure its specificity. Finally, the function $PCRlen_{ratio}(P_v)$ calculates the appropriate lengths of the PCR products.

Algorithm

The proposed algorithm consists of six processes: 1) random population initialization, 2) local search, 3) fitness evaluation, 4) selection, crossover, and mutation operations, 5) replacement operation, and 6) judgment on termination conditions. The six processes are described below:

1) Random population initialization. To start the algorithm, chromosomes $P_v = (F_{l1}, P_{l1}, R_{l1}, F_{l2}, P_{l2}, R_{l2})$ of particular number are randomly generated for an initial population without duplicates. F_{l1}, R_{l1}, F_{l2} and R_{l2} are randomly generated between the minimum length and the maximum length given by the primer length constraint. The minimum and maximum lengths are set to 16 and 28 bp, respectively. The PCR product lengths, P_{l1} and P_{l2} are randomly generated between 100 bp and δ_1, and between 100 bp and δ_2, respectively. (δ_1 and δ_2 are the maximum tolerant PCR product lengths of P_{l1} and P_{l2} shown in Fig.1)

2) Local search. The local search identifies superior individuals from amongst the neighbors of the original individual. The experience of the original individual is improved, and thus a local optimum solution can be obtained. At each iteration, all offsprings are subjected to the local search process so that local optimality is always preserved. Finally, a global optimality is determined. The pseudo-code for the local search is shown below:

Local search pseudo-code

```
1  Begin;
2      Select an incremental value d=a*Rand();
3      For a given chromosome i ∈ P: calculate fitness (i);
4      For j=1 to number of variables in chromosome i;
5          value(j)= value(j)+d;
6      If chromosome fitness does not improve then
7          value(j)= value(j)-d;
8      else if chromosome fitness improves then
9          retain value(j);
10     Next j;
11 End;
```

In the local search, P represents a population and d is calculated to assist an individual in seeking out neighboring individuals. Before calculating d, a must be determined, which is a constant that suits the variable values. We selected the variable values of vector P_v based on a maximum permitted flexible range. F_{l1}, R_{l1}, F_{l2} and R_{l2} are selected to determine the constant a during implementation. Consequently, the constant a is the difference of the maximum primer length and the minimum primer length and was set to 12 bps. In the proposed method, it only searches in a straight line in the six-dimensional space. This saves a considerable amount of processing time.

3) Fitness evaluation. The fitness value in the fitness function is used to individually ascertain whether a chromosome is good or bad. We use formula (3) to evaluate the fitness values of all chromosomes in the population for related chromosomal operations later.

4) Selection, crossover, and mutation operations. In MAs, the processes for evolutionary computation include selection, crossover and mutation. A random selection is applied to select two chromosomes from the population. The selected two chromosomes are processed by the crossover operation. Uniform crossover is used to implement the crossover operation. One-point mutation is applied in the proposed MA.

5) Replacement operation. After the evolutionary computation processes have been implemented, the two worst chromosomes in a population are replaced by the new offsprings, and the process is repeated in the next generation.

6) Judgment on termination conditions. Once an optimal solution of chromosomes or the maximum

number of iterations is reached or the fitness value is 0, the MA is terminated.

3 Results

The *in silico* environment

The proposed method was run on a Xeon(TM) CPU 3.20 GHz × 2 and 2GB RAM under the Microsoft Windows XP SP2 and JAVA 5.0 platforms.

Template sequence

A point mutation in the SLC6A4 gene was recently identified and shown to be associated with autism spectrum disorders [16], psychosis [17], and bipolarity [18] in patients. The SLC6A4 gene is the member 4 for a solute carrier family 6 (neurotransmitter transporter, serotonin). Overall, 288 SNPs in the SLC6A4 gene were used to evaluate the efficiency of the proposed method. All template sequences were retrieved with a 500 bp flanking length (at both sides of SNP) from SNP-Flankplus (http://bio.kuas.edu.tw/snp-flankplus/) [19].

Parameter settings

Four parameters need to be set for the proposed algorithm and the GA, i.e. the number of iterations (generations), the population size, the probability of crossover and the probability of mutation. The respective values were 1000, 50, 0.6 and 0.001. These values are based on DeJong and Spears' parameter settings [20]. Furthermore, constraints commonly used in CTPP primer design, such as a primer length of between 16~28 bp, a GC% between 20~80%, a primer T_m between 45 and 62°C, a difference of CTPP primer T_m of less than 1°C, and a product length larger than 100 bp were used. Finally, the PCR product length was set to three ratios (ratios 1, 2, and 3) at 8, 13, and 20, respectively. This allowed for the distinct separation of PCR bands via gel electrophoresis. These ratios were chosen based on PCR experiments previously conducted by us [13].

Results of the GA-based method

The CTPP primer set results obtained by the GA-based CTPP primer design method with the common constraints are shown in Table 1. For the 288 SNPs, the primer lengths all lay between 16 and 28 bp. Table 1 shows that for the parameter settings of DeJong and Spears, 75.12% of the designed primers satisfy the length difference criterion. Most of the primer length differences were between 0 and 5 bp (data not shown). Of the primers, 96.09% primers satisfied GC% the criterion; only 30 primers had a GC% of less than 20%, and 25 primers had one higher than 80% (data not shown). The GC clamp criterion was satisfied by 55.99% of the primers. Most of the designed primers also satisfied the T_m criterion (86.63%). However, only a few of the primer pairs satisfied the T_m difference criterion (23.61%). The product length criterion was complied with by 71.18% of the designed primer pairs. And finally, only a few primers (4.44%, 14.06% and 3.04%, respectively) did not satisfy the dimer, hairpin and specificity criteria.

Results for the proposed method

The CTPP primer set results obtained by the MA-based CTPP primer design method with the common constraints are shown in Table 1. For the 288 SNPs, the primer lengths all lay between 16 and 28 bp. Table 1 shows that for the parameter settings of DeJong and Spears, 84.95% of the designed primers satisfy the length difference criterion. Most of the primer length differences were between 0 and 5 bp (data not shown). Of the primers, 96.70% primers satisfied GC% the criterion; only 24 primers had a GC% of less than 20%, and 19 primers had one higher than 80% (data not shown). The GC clamp criterion was satisfied by 60.50% of the primers. Most of the designed primers also satisfied the T_m criterion (96.53%). However, only a few of the primer pairs satisfied the T_m difference criterion (60.88%). The product length criterion was complied with by 68.98% of the designed primer pairs. And finally, only a few primers (3.96%, 17.27% and 1.91%, respectively) did not satisfy the dimer, hairpin and specificity criteria.

Table 1. Compliance (%) of the GA-based and MA-based methods with the DeJong and Spears parameter settings with the primer design contraints for SNPs of the SLC6A4 gene.

	Constraints								
Method	primer length difference	GC%	GC clamp	T_m	T_m difference	product length	dimer	hairpin	specificity
GA	75.12	96.09	55.99	86.63	23.61	71.18	95.56	85.94	96.96
MA	84.95	96.70	60.50	96.53	60.88	68.98	96.04	82.73	98.09

Comparison of the MA-based and GA-based results

As Table 1 shows most compliances of the MA-based method is better than the GA-based method. The compliance with the primer length difference criterion of the MA-based method is 9.83% higher than for the GA-based method. The compliance with the GC% criterion is slightly higher (0.61%) for the MA-based method than for the GA-based method. The GC clamp compliance of the MA-based method is 4.51% higher than the one of the GA-based method. The T_m and T_m difference compliances of the MA-based method are 9.90% and 37.27% higher than for the GA-based method. And finally, compliances with the dimer and specificity criteria of the MA-based method are 0.48% and 1.13% higher than for the GA-based method. However, the product length and hairpin compliances are lightly 2.20% and 3.21% lower than for the GA-based method.

4 Discussion

In this study, we developed a MA-based algorithm which has been shown to be a robust search and optimization method in a number of practical applications, especially for problems with highly complex SNP genotyping with the CTPP primer design functions. *In silico* simulations validated the reliability of the MA-based CTPP primer design method.

Influence of annealing temperatures

In PCR-CTPP, the designed annealing temperatures of the primers are more important than in PCR-RFLP. When the T_m value is similar among the four PCR-CTPP primers, the PCR competition between all possible DNA products is balanced [12]. When the annealing temperature is low, the PCR reactions are non-specific, leading to incorrect heterozygous genotyping. Therefore, a non-competitive or specific amplification is important to correctly genotype SNPs using PCR-CTPP. This problem is resolved by computationally finding similar T_m values for the four CTPP primers and by experimentally adjusting the annealing temperature for the PCR experiment [12, 21]. The MA used in this study to design the PCR-CTPP primers improves the efficiency by finding almost identical T_m values for the four primers. By conforming to the T_m difference criterion in the *in silico* simulations (Table 1), the compliance of the MA-based method is considerably raised (37.27%) compared to the GA-based method [13]. This shows that the MA-based method more reliably fits the T_m constraint to the primers than the previously developed GA-based method.

Primer design constraints concerned

Since the T_m is important in the proposed MA-based PCR-CTPP primer design method, it could be assumed that other factors might also improve this automated system. Further basic research is required to determine these factors. This study investigates only the typical primer design constraints, such as primer length, difference of primer pair length, GC proportion, PCR product length, GC clamp, dimer of primer pair (including cross-dimers and self-dimers), hairpin, PCR product length and specificity.

5 Conclusion

The MA-based CTPP primer design method provides critical melting temperature and all kind of common primer constraints estimation. The MA-based method yields more feasible CTTP primer set than GA-based method. The experimental flexibility of the MA designed PCR-CTPP primers for 288 polymorphisms have been confirmed by *in silico* simulations. In future, PCR-CTPP may replace PCR-RFLP because the restriction enzyme digestion step can be skipped. This is associated with a lower cost and shorter genotyping times [12]. In conclusion, the proposed MA-based method is ideally suited for the design of feasible CTPP primers since it conforms to most of the PCR-CTPP constraints.

Acknowledgements

This work is partly supported by the National Science Council in Taiwan under grant 98-2221-E-151-040-, 98-2622-E-151-001-CC2 and 98-2622-E-151-024-CC3.

References

[1] F. Jasmine, H. Ahsan, I. L. Andrulis, E. M. John, J. Chang-Claude, and M. G. Kibriya, "Whole-genome amplification enables accurate genotyping for microarray-based high-density single nucleotide polymorphism array," *Cancer Epidemiol Biomarkers Prev,* vol. 17, pp. 3499-508, Dec 2008.

[2] L. Hui, T. DelMonte, and K. Ranade, "Genotyping using the TaqMan assay," *Curr Protoc Hum Genet,* vol. Chapter 2, p. Unit 2 10, Jan 2008.

[3] 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, p. 30, 2006.

[4] G. T. Lin, H. F. Tseng, C. H. Yang, M. F. Hou, L. Y. Chuang, H. T. Tai, M. H. Tai, Y.

H. Cheng, C. H. Wen, C. S. Liu, C. J. Huang, C. L. Wang, and H. W. Chang, "Combinational polymorphisms of seven CXCL12-related genes are protective against breast cancer in Taiwan," *OMICS,* vol. 13, pp. 165-172, Jan 23 2009.

[5] T. Aomori, K. Yamamoto, A. Oguchi-Katayama, Y. Kawai, T. Ishidao, Y. Mitani, Y. Kogo, A. Lezhava, Y. Fujita, K. Obayashi, K. Nakamura, H. Kohnke, M. Wadelius, L. Ekstrom, C. Skogastierna, A. Rane, M. Kurabayashi, M. Murakami, P. E. Cizdziel, Y. Hayashizaki, and R. Horiuchi, "Rapid single-nucleotide polymorphism detection of cytochrome P450 (CYP2C9) and vitamin K epoxide reductase (VKORC1) genes for the warfarin dose adjustment by the SMart-amplification process version 2," *Clin Chem,* vol. 55, pp. 804-12, Apr 2009.

[6] 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 Res,* vol. 28, pp. 2001-7, Jul-Aug 2008.

[7] N. Hamajima, T. Saito, K. Matsuo, K. Kozaki, T. Takahashi, and K. Tajima, "Polymerase chain reaction with confronting two-pair primers for polymorphism genotyping," *Jpn J Cancer Res,* vol. 91, pp. 865-8, Sep 2000.

[8] N. Katsuda, N. Hamajima, A. Tamakoshi, K. Wakai, K. Matsuo, T. Saito, K. Tajima, and S. Tominaga, "Helicobacter pylori seropositivity and the myeloperoxidase G-463A polymorphism in combination with interleukin-1B C-31T in Japanese health checkup examinees," *Jpn J Clin Oncol,* vol. 33, pp. 192-7, Apr 2003.

[9] S. Togawa, T. Joh, M. Itoh, N. Katsuda, H. Ito, K. Matsuo, K. Tajima, and N. Hamajima, "Interleukin-2 gene polymorphisms associated with increased risk of gastric atrophy from Helicobacter pylori infection," *Helicobacter,* vol. 10, pp. 172-8, Jun 2005.

[10] K. K. Abu-Amero, O. M. Al-Boudari, G. H. Mohamed, and N. Dzimiri, "The Glu27 genotypes of the beta2-adrenergic receptor are predictors for severe coronary artery disease," *BMC Med Genet,* vol. 7, p. 31, 2006.

[11] S. J. Yang, H. Y. Wang, X. Q. Li, H. Z. Du, C. J. Zheng, H. G. Chen, X. Y. Mu, and C. X. Yang, "Genetic polymorphisms of ADH2 and ALDH2 association with esophageal cancer risk in southwest China," *World J Gastroenterol,* vol. 13, pp. 5760-4, Nov 21 2007.

[12] N. Hamajima, T. Saito, K. Matsuo, and K. Tajima, "Competitive amplification and unspecific amplification in polymerase chain reaction with confronting two-pair primers," *J Mol Diagn,* vol. 4, pp. 103-7, May 2002.

[13] C. H. Yang, Y. H. Cheng, L. Y. Chuang, and H. W. Chang, "Genetic Algorithm for the Design of Confronting Two-Pair Primers," *Ninth IEEE international Conference on BioInformatics and BioEngineering (BIBE),* pp. 242-247, 2009.

[14] P. Moscato, "On evolution, search, optimization, genetic algorithms and martial arts: Towards memetic algorithms," Tech. Rep. Caltech Concurrent Computation Program, Rep. 826, California Inst. Technol., Pasadena, CA 1989.

[15] R. Dawkins, *The Selfish Gene*: Oxford University Press, USA, 2006.

[16] 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 Res,* vol. 1, pp. 251-7, Aug 2008.

[17] 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 Res,* vol. 111, pp. 103-8, Jun 2009.

[18] L. Mandelli, M. Mazza, G. Martinotti, M. Di Nicola, T. Daniela, E. Colombo, S. Missaglia, D. De Ronchi, R. Colombo, L. Janiri, and A. Serretti, "Harm avoidance moderates the influence of serotonin transporter gene variants on treatment outcome in bipolar patients," *J Affect Disord,* p. Epub ahead of print, Mar 28 2009.

[19] 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. 147-9, 2008.

[20] K. A. De Jong and W. M. Spears, "An analysis of the interacting roles of population size and crossover in genetic algorithms." vol. 1: Springer, 1990, pp. 38–47.

[21] N. Hamajima, "PCR-CTPP: a new genotyping technique in the era of genetic epidemiology," *Expert Rev Mol Diagn,* vol. 1, pp. 119-23, May 2001.