Multiplex PCR Assay Design by Hybrid Multiobjective Evolutionary Algorithm

In-Hee Lee, Soo-Yong Shin, and Byoung-Tak Zhang

Biointelligence Laboratory
School of Computer Science and Engineering
Seoul National University, Seoul 151-742, Korea
{ihlee,syshin,btzhang}@bi.snu.ac.kr

Abstract. Multiplex Polymerase Chain Reaction (PCR) assay is to amplify multiple target DNAs simultaneously using different primer pairs for each target DNA. Recently, it is widely used for various biology applications such as genotyping. For sucessful experiments, both the primer pairs for each target DNA and grouping of targets to be actually amplified in one tube should be optimized. This involves multiple conflicting objectives such as minimizing the interaction of primers in a group and minimizing the number of groups required for the assay. Therefore, a multiobjective evolutionary approach may be an appropriate approach. In this paper, a hybrid multiobjective evolutionary algorithm which combines ϵ-multiobjective evolutionary algorithm with local search is proposed for multiplex PCR assay design. The proposed approach was compared with another multiobjective method, called MuPlex, and showed comparative performance by covering all of the given target sequences.

1 Introduction

The Polymerase Chain Reaction (PCR) is a very powerful biological technique which is widely used to amplify DNA and plays a key role in biotechnology and biology research. In standard protocol, PCR can amplify only one target DNA at a time (Fig. 1(a)). But the biological or clinical assay usually involves multiple target DNAs, it is much more desirable to amplify these DNAs simultaneously. The multiplex PCR is an extension of PCR in which multiple target DNAs are amplified at the same time (Fig. 1(b)). It has a wide variety of applications in biology and is recently spotlighted as a core tool for high throughput single nucleotide polymorphism (SNP) genotyping [1,2,3].

For successful experimental results, a careful design of multiplex PCR assay is important. A multiplex PCR assay design is a complex problem composed of two optimization processes: optimizing primers for each target while minimizing the number of partition. First, the primers for each target DNA should be optimized so that the interactions between primers and non-target DNAs are minimized. Since the multiple targets are amplified in one tube at the same time, it is important that the primers for one target do not interact another targets or primers. If such an unwanted interaction happens, some of the target DNAs

S. Obayashi et al. (Eds.): EMO 2007, LNCS 4403, pp. 376–385, 2007.

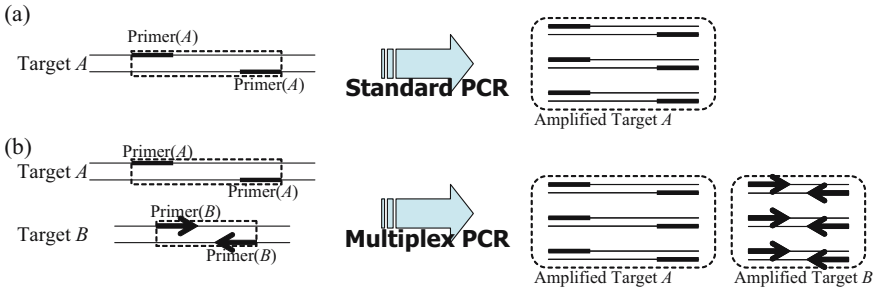

Fig. 1. (a) The concept of standard polymerase chain reaction (PCR). The region between primers in A (the dashed box) is amplified by PCR. (b) The concept of multiplex PCR. Multiple targets A and B are amplified simultaneously by primers A and B, respectively, in one experiment.

might not be amplified. Last, the grouping of target DNAs which will be amplified together should be decided. It would be ideal when all of the target DNAs can be amplified together. Unfortunately, it is very likely that the primers for some targets can not be chosen to prevent unwanted interactions. In such cases, the targets should be put to different groups for separate multiplex PCR runs. However, the number of such separate groups should be minimized.

There have been many studies to tackle this problem [4,5,6,7,8,9,10,11,12] . Most of these works assumed only one group and the targets that do not fit to be amplified together were discarded [5,6,7,8,11,12]. In [4,9,10], on the other hand, the partitioning of targets into multiple groups was handled. First, a set of primer candidates for each target is selected according to predefined conditions. Then, the targets are partitioned into appropriate groups in deterministic way while selecting optimal primers from the candidate sets. From the methodological point of view, most of the previous researches used a deterministic search and only a few evolutionary approaches are published [7,8,11].

Rachlin et al. formulated the design of multiplex PCR assay as finding cliques in graph to optimize both objectives [13]. According to their formulation, the nodes in graph G represent the target DNAs and edges connect two targets(nodes) which can be put into the same group. Each node has multiple states (candidate primers) and the state of two nodes determines whether they can be connected or not. They empirically showed that there is a tradeoff relationship between the specificity of each primer pair to their target and the overall degree of multiplexing. Moreover, it is well known that finding a clique in a given graph is a hard computational problem [14]. Considering these properties of multiplex PCR assay design, a multiobjective evolutionary approach with local search is suggested here.

The suggested multiobjective evolutionary algorithm is based on ϵ-MOEA which was originally suggested in [15]. The algorithm is modified to perform local search after the generation of every new offspring and a genotypical niching is adopted to keep the population diversity.

The rest of the paper is organized as follows. Section 2 describes the multiplex PCR assay design and the strategy taken here in more detail. The hybrid multiobjective evolutionary algorithm used in multiplex PCR assay design is explained in Section 3. The experimental results are shown and compared to other methods in Section 4 and the conclusions are drawn in Section 5.

2 Multiplex PCR Assay Design

As already mentioned in Section 1, two problems should be solved for a successful multiplex PCR assay. One is to select primers for each target. Those primers must have uniform experimental conditions and minimize the interaction with other targets and primers. The other is to divide the set of targets into multiple subsets. For example, some target DNAs share similar subsequences. In that case, it is very likely that one target's primer can interact with other target and vice versa. Therefore, these targets should be separated into different subsets. But the number of different subsets needs to be minimized to simplify the experimental process. However, we can not put all compatible targets together because there is a limitation on the number of targets that can be amplified in one tube for technical reasons. Hence we should minimize the number of different subsets while keeping the maximum size of each subset.

The primers having uniform experimental conditions can be chosen independently for each target. But the minimization of the interaction between other targets and primers depends on the partitioning of targets. The partitioning of targets is also dependent on the selection of primers. Therefore, the selection of primer candidates can be separated from the optimization process. From now on, we assume that a set of primer candidates for each target is given *a priori* and concentrate on primer assignment on each target from the candidates and partition of targets.

In [13], it is more formally described using the concept of multi-node graph. In a multi-node graph, each node has its own set of states it can take. In multiplex PCR context, a node corresponds to a target and a node's state means the primers for the target. If a node u is in state i, the i-th primer candidate is assigned for target u. Each edge in the graph is associated with variable weight depending on the states on the two nodes it connects (see Fig. 2). We denote the

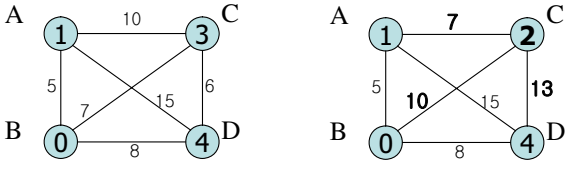

Fig. 2. An example of multi-node graph (left). After changing the state of node C, the weights on edges connected to C are changed as the second graph (right).

weight values for edge between nodes u and v by the matrix W_{uv}. For multiplex PCR assay design, the weight matrix denotes the compatibility between two targets, which means whether the targets can be put together in one tube. The elements of $W_{uv}[i][j]$ represents how compatible the targets u and v are when they are in states i and j respectively. For the two targets to be compatible, they should satisfy two conditions: minimizing the undesirable hybridization and the sequence similarity among targets and their primers. Minimizing the undesirable hybridization alone is not enough because similar sequences can also reduce the hybridization chance. For example, the sequences 'AAAA' and 'CCCC' do not hybridize each other, but 'AAAA' and 'AAAA' do not, either. Hence, we decompose the compatibility between targets (W_{uv}) by two values: H-measure (H_{uv}) and Similarity (S_{uv}) defined in [16]. H_{uv} is a matrix whose element $H_{uv}[i][j]$ denotes how much undesirable hybridization can occur among targets u and v and their primers i and j. Similarly, S_{uv} is a matrix whose element $S_{uv}[i][j]$ denotes how similar the targets u and v and their primers i and j are.

Given a set of n targets T and the set of primer candidates C_i for each target i, a multi-node graph G and its associated matrices H and S can be constructed. Then, formally, the multiplex PCR assay design is to find,

1. The partition $S_1, \ldots, S_M$ such that $\bigcup S_i = T$ and $S_i \cap S_j = \varnothing$ for each i, j and
2. The state assignment A for each node in G from $C_1 \times C_2 \times \cdots \times C_n$,

while satisfying the following:

1. Minimize $\sum_i \sum_{u,v \in S_i} H_{uv}[A(u)][A(v)]$ and
2. Minimize $\sum_i \sum_{u,v \in S_i} S_{uv}[A(u)][A(v)]$ and
3. Minimize the number of partitions M,

where $A(u)$ denotes the state of node u under assignment A.

According to the above definition, two different variable spaces should be searched: state assignment and partition. We explored both space by combining a multiobjective evolutionary algorithm and local search. Here, the main evolutionary algorithm searches the space of the state assignment. During local search, the space of partition is explored. Detailed description of the evolutionary search procedure will be given in the next section.

Our approach to multiplex PCR assay design is summarized in Fig. 3. First, candidate primers are generated by Primer3 [17]. We used an external program at this step from two reasons: one is that the candidate primers can be selected independently from the main evolutionary optimization and the other is that there exist many open softwares for primer selection since it is a fundamental tool in biology. Among various open softwares, we chose the most popular program, Primer3. Next, the partition of targets and primer assignment for each target are optimized by hybrid multiobjective evolutionary algorithm. At the end, a variety of multiplex PCR assays will be presented to user.

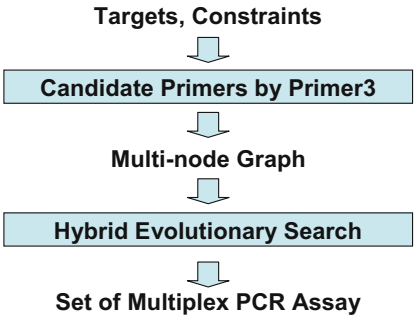

Fig. 3. The flow chart for Multiplex PCR assay design

3 Hybrid Multiobjective Evolutionary Algorithm for Multiplex PCR Assay Design

3.1 The Preprocessing

The first step is to select primer candidates for each target which having similar chemical properties. We used Primer3 program to choose candidate primers [17]. Considering the size of the search space and running time, five candidate primers are selected for each target. And for each pair of target u and v, the elements of $H_{uv}[i][j]$ and $S_{uv}[i][j]$ are calculated using the H-measure and Similarity function described in [16].

3.2 The Hybrid Multiobjective Evolutionary Algorithm

Our next step is the evolutionary search for optimal partition of groups and primer assignments. We used a variation of ϵ-MOEA [15] combined with local search.

Each individual is a concatenation of partition part and state assignment part. The state assignment is a vector from $C_1 \times C_2 \times \cdots \times C_n$ which determines the configuration of multi-node graph. The i-th value denotes which primer candidate from C_i is assigned for target i. The partition part means the set of cliques from the multi-node graph configured by the state assignment part. It is a vector from $[1, \ldots, M]^n$ where i-th value determines which partition the node i belongs to.

For each generation of ϵ-MOEA,

1. One parent P_1 is chosen at random from the archive and the other parent P_2 is chosen from the population by tournament selection.
2. Generate two offsprings O_1 and O_2 from P_1 and P_2 by genetic operators. The uniform crossover and 1-bit mutation operators are used here.
3. Apply the local search to O_1 and O_2 for a predefined number of times, L.

   (a) O'_1 and O'_2 is produced through local search operators. One of the two local search operators which will be explained in Section 3.2 is applied in random.
   (b) Replace O_1 with O'_1 if O'_1 dominates O_1. If O_1 and O'_1 non-dominates each other, O'_1 replaces O_1 with probability of 0.5. Otherwise, O'_1 is discarded.
   (c) Similar procedure with O_2 and O'_2.
4. Update the archive.
   (a) The offspring is accepted to the archive if it is in the same front as the archive members or it dominates one or more of them. The dominated archive members are removed.
   (b) If the archive reaches its maximum size, the nearest archive member in objective space is replaced.
5. Update the population.
   (a) If the offspring dominates one of the population, the dominated member is replaced with the offspring.
   (b) If the offspring and the population do not dominate each other, the nearest population member in variable space is replaced with probability of 0.5.
6. Repeat to Step 1 until the termination condition is satisfied.

In original ϵ-MOEA, Step 4 followed after Step 2 and there was no restriction on maximum size of the archive. Theoretically, the number of archive members that ϵ-dominate the population is not infinite [15]. However, the size of the archive often grows very large. So we put a limit on the maximum size of the archive.

Another difference between the suggested approach and the original ϵ-MOEA is the niching method. In original ϵ-MOEA, there was no explicit niching method for the archive except the ϵ-domination concept and the new offsprings replaced random individual from the population. However, the ϵ-domination concept is not enough to keep the size of the archive in reasonable size and the random replacement in the population can lead to the quick loss of genetic diversity. As can be seen from Step 4 and 5, we tried to handle the problem by using different distance measures. In archive, the distance in objective space is used to provide diverse solutions. In contrast, the distance in variable space is used in population update. This is to keep the genetic diversity of population and to prevent premature convergence.

3.3 Local Search

When generating new offsprings in Step 2 of the main loop, the crossover and mutation operator targets the entire chromosome. Both the state assignment vector and the partition vector are treated as one string of size $2n$ and undergo uniform crossover or 1-bit mutation.

On the other hand, the local search operator targets the partition vector only. Since every candidate primers generated by Primer3 guarantee minimum level

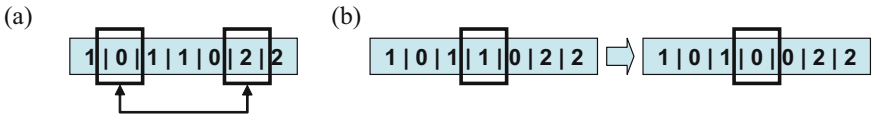

Fig. 4. The two local search operators. (a) The swapping operator exchanges two targets between two different partitions. (b) The migration operator moves one target from a partition to another.

of quality, the change of primers assigned to a target does not result in drastic change in objective values. Hence, we concentrate on finding optimal partition during local search. As can be seen in Fig. 4, two local search operators are adopted. One is the swapping operator that exchanges two targets from two different partitions. The other is the migration operator which moves one target from a partition to another.

Local search proceeds in Lamarckian way. At each cycle of local search, the individual produced by local search which dominates the previous replaces the original individual. If the individuals before and after the local search non-dominates each other, one is chosen at random.

4 Experimental Results

We tested our approach on 52 target sequences from Arabidopsis multigene family and compared the result with an open program for multiplex PCR assay design, MuPlex [10]. Among the MULTIPCR [4], MultiPLX [9] and MuPlex [10] that can handle multiple partition, MuPlex was chosen because MULTIPCR was not open to public and MultiPLX could not find any acceptable result for the given problem.

MuPlex uses an agent-based multi-objective optimization. The agents encapsulating specific algorithm either create new solutions from scratch, improve or modify existing solutions, or remove unpromising solutions from further consideration [10]. By the interactions between agents, MuPlex implements similar approach as evolutionary algorithm. The solutions in MuPlex are evaluated in similar terms as our approach.

We set the size of population and archive as 100 and 200, respectively. The maximum generation was set to 100,000 and local search was performed 100 times for each offspring. The probability for crossover and mutation was 0.9 and 0.01, respectively. The maximum number of partition and maximum size of a partition is set to 10. These experimental parameters were chosen empirically.

The results are evaluated from three perspectives. One is the sum of total cross-hybridization within a partition. This is to estimate total experimental errors. The DNA-DNA hybridization simulator NACST/Sim is used to calculate this value [16]. It checks all possibilities of cross-hybridization between two given sequences. Others are the number of groups and the average number of targets per group. These are to estimate the efficiency of multiplex PCR assay.

Table 1. The comparison of designed Multiplex PCR Assay. Solution 1, Solution 2 and Solution 3 are generated by the hybrid multiobjective evolutionary algorithm suggested in this paper. The solution named MuPlex is generated using MuPlex program.

	MuPlex	Solution 1	Solution 2	Solution 3
The total cross-hybridization	13719	10915	13269	10683
The number of groups	5	9	8	10
The average number of targets	9.4	5.7	6.5	5.2

Table 2. The design examples from MuPlex and the proposed approach. The Solution 2 from Table 1 is shown here. The columns group ID and group size denote each partition of target DNAs and the number of targets in each partition, respectively. # of cross-hyb. value means the total undesirable hybridization value calculated by NACST/Sim.

MuPlex			Hybrid ϵ-MOEA		
Group ID	Group Size	# of cross-hyb.	Group ID	Group Size	# of cross-hyb.
1	10	4926	1	6	1746
2	10	2397	2	7	1385
3	10	2789	3	6	1655
4	10	2032	4	6	2295
5	7	1575	5	6	1234
			6	7	1493
			7	6	1912
			8	8	1549
Total	47	13719	Total	52	13269

The best run of our algorithm output only three solutions in the final archive. These solutions are compared with the only solution from MuPlex in Table 1. From the three solutions produced by our approach, the tradeoff between primer optimization and partition efficiency is clear. As the number of group increases, the average number of targets in each group decreases. And if the number of targets in a group is small, there is little chance to the cross-hybridization. In that sense, all of the four solutions in Table 1 form a tradeoff front.

The design examples from MuPlex and the proposed approach are compared in detail in Table 2. The columns group ID and group size denote each partition of target DNAs and the number of targets in each partition, respectively. NACST/Sim value means the total undesirable hybridization value calculated by NACST/Sim. In MuPlex, some target can be discarded if it is hard to find a partition for that target. Therefore, as can be seen in Table 2, only 47 of 52 targets were partitioned. In contrast, every target belongs to a partition in our approach. But this is dependent on the purpose of the user. In some case as high throughput screening, users need a design which is efficient but do not cover every target. But in cases of clinical assay, the coverage becomes critical. Also, the constraint of perfect coverage upon the proposed approach can be relaxed.

5 Conclusions

The problem of multiplex PCR assay design is addressed and formulated as a multiobjective optimization problem. A hybrid multiobjective evolutionary search is applied to the problem and compared with other similar program. The suggested approach combines a variant of existing algorithm and two simple local search operators and shows a reasonable performance. This is a preliminary result and further work is required.

Acknowledgements

This research was supported in part by the Ministry of Education & Human Resources Development under the BK21-IT Program, the Ministry of Commerce, Industry and Energy through MEC project, and the NRL Program from Korean Ministry of Science and Technology. The ICT at Seoul National University provides research facilities for this study. Soo-Yong Shin was supported by the Korea Research Foundation Grant funded by the Korean Government (MOEHRD) (KRF-2006-214-D00140).

References

1. A. L. L. Cortez, A. C. Carvalho, A. A. Ikuno, K. P. Bürger, and A. M. C. Vidal-Martins. Identification of *Salmonella* spp. isolates from chicken abattoirs by multiplex-PCR. *Research in Veterinary Science*, 81(3):340–344, 2006.
2. Christina L. Aquilante, Taimour Y. Langaee, Peter L. Anderson, Issam Zineh, and Courtney V. Fletcher. Multiplex PCR-pyrosequencing assay for genotyping CYP3A5 polymorphisms. *Clinica Chimica Acta*, 372(1–2):195–198, 2006.
3. Anne J. Jääskeläinen, Heli Piiparinen, Marija Lappalainen, Marjaleena Koskiniemi, and Vaheri Antti. Multiplex-PCR and oligonucleotide microarray for detection of eight different herpesviruses from clinical specimens. *Journal of Clinical Virology*, 37(2):83–90, 2006.
4. Pierre Nicodème and Jean-Marc Steyaert. Selecting optimal oligonucleotide primers for multiplex PCR. In *Proceedings of the 5th International Conference on Intelligent Systems for Molecular Biology*, pages 210–213, 1997.
5. Eric C. Rouchka, Abdelnaby Khalyfa, and Nigel G. F. Cooper. MPrime: efficient large scale multiple primer and oligonucleotide design for customized gene microarrays. *BMC Bioinformatics*, 6(175), 2005.
6. Richard Schoske, Pete M. Vallone, Christian M. Ruitberg, and John M. Butler. Multiplex PCR design strategy used for the simultaneous amplification of 10 Y chromosome short tandem repeat (STR) loci. *Analytical and Bioanalytical Chemistry*, 375:333–343, 2003.
7. Hong-Long Liang, Chungnan Lee, and Jain-Shing Wu. Multiplex PCR primer design for gene family using genetic algorithm. In *GECCO '05*, pages 67–74, Washington, DC, USA, 2005. ACM.
8. Feng-Mao Lin, Hsien-Da Huang, His-Yuan Huang, and Jorng-Tzong Horng. Primer design for multiplex PCR using a genetic algorithm. In *GECCO '05*, pages 475–476, Washington, DC, USA, 2005. ACM.

9. Lauris Kaplinski, Peidar Andreson, Tarmo Puurand, and Maido Remm. MultiPLX: automatic grouping and evaluation of PCR primers. *Bioinformatics*, 21(8):1701–1702, 2005.
10. John Rachlin, Chunming Ding, Charles Cantor, and Simon Kasif. MuPlex: multi-objective multiplex PCR assay design. *Nucleic Acids Research*, 33:W544–W547, 2005.
11. Chungnan Lee, Jain-Shing Wu, Yow-Ling Shiue, and Hong-Long Liang. Multi-Primer: Software for multiple primer design. *Applied Bioinformatics*, 5(2):99–109, 2006.
12. Tomoyuki Yamada, Haruhiko Soma, and Shinichi Morishita. PrimerStation: a highly specific multiplex genomic PCR primer design server for the human genome. *Nucleic Acids Research*, 34:W665–W669, 2006.
13. John Rachlin, Chunming Ding, Charles Cantor, and Simon Kasif. Computational tradeoffs in multiplex PCR assay design for SNP genotyping. *BMC Genomics*, 6(102), 2005.
14. M. R. Garey and D. S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-completeness*. W. H. Freeman and company, 1979.
15. Marco Laumanns, Lothar Thiele, Kalyanmoy Deb, and Eckart Zitzler. Combining convergence and diversity in evolutionary multi-objective optimization. *Evolutionary Computation*, 10(3):263–282, 2002.
16. Soo-Yong Shin, In-Hee Lee, and Byoung-Tak Zhang. Multi-objective evolutionary optimization of DNA sequences for reliable DNA computing. *IEEE Transactions on Evolutionary Computation*, 9(2):143–158, 2005.
17. Steve Rozen and Helen J. Skaletsky. Primer3 on the www for general users and for biologist programmers. In S. Krawetz and S. Misener, editors, *Bioinformatics Methods and Protocols: Methods in Molecular Biology*, pages 365–386. Humana Press, Totowa, NJ, 2000. Source code available at http://fokker.wi.mit.edu/primer3/.