BIOINFORMATICS ORIGINAL PAPER Vol. 21 no. 24 2005, pages 4330–4337
doi:10.1093/bioinformatics/bti730

*Sequence analysis*

# Integrated minimum-set primers and unique probe design algorithms for differential detection on symptom-related pathogens

Yu-Cheng Huang[1,†], Chun-Fan Chang[3,†], Chen-hsiung Chan[1], Tze-Jung Yeh[1,2], Ya-Chun Chang[2], Chaur-Chin Chen[4] and Cheng-Yan Kao[1,5,*]

[1]Bioinformatics Laboratory, Department of Computer Science and Information Engineering and [2]Department of Plant Pathology and Microbiology, National Taiwan University, Taipei, Taiwan, [3]Breed-Use-Special Laboratory, Center of Agriculture Hierarchical Utilization, Graduate Institute of Biotechnology, Chinese Culture University, Taipei, Taiwan, [4]Department of Computer Science, National Tsing-Hua University, Hsinchu, Taiwan and [5]Institute for Information Industry, Taipei, Taiwan

Received on May 27, 2005; revised on October 14, 2005; accepted on October 18, 2005
Advance Access publication October 25, 2005

**ABSTRACT**

**Motivation:** Differential detection on symptom-related pathogens (SRP) is critical for fast identification and accurate control against epidemic diseases. Conventional polymerase chain reaction (PCR) requires a large number of unique primers to amplify selected SRP target sequences. With multiple-use primers (mu-primers), multiple targets can be amplified and detected in one PCR experiment under standard reaction condition and reduced detection complexity. However, the time complexity of designing mu-primers with the best heuristic method available is too vast. We have formulated minimum-set mu-primer design problem as a set covering problem (SCP), and used modified compact genetic algorithm (MCGA) to solve this problem optimally and efficiently. We have also proposed new strategies of primer/probe design algorithm (PDA) on combining both minimum-set (MS) mu-primers and unique (UniQ) probes. Designed primer/probe set by PDA-MS/UniQ can amplify multiple genes simultaneously upon physical presence with minimum-set mu-primer amplification (MMA) before intended differential detection with probes-array hybridization (PAH) on the selected target set of SRP.

**Results:** The proposed PDA-MS/UniQ method pursues a much smaller number of primers set compared with conventional PCR. In the simulation experiment for amplifying 12 669 target sequences, the performance of our method with 68% reduction on required mu-primers number seems to be superior to the compared heuristic approaches in both computation efficiency and reduction percentage. Our integrated PDA-MS/UniQ method is applied to the differential detection on 9 plant viruses from 4 genera with MMA and PAH of 11 mu-primers instead of 18 unique ones in conventional PCR while amplifying overall 9 target sequences. The results of wet lab experiments with integrated MMA-PAH system have successfully validated the specificity and sensitivity of the primers/probes designed with our integrated PDA-MS/UniQ method.

**Contact:** cykao@csie.ntu.edu.tw

**Supplementary information:** http://www.csie.ntu.edu.tw/~cykao/pda/

## 1 INTRODUCTION

Optimal design of DNA primers and probes aids the diagnosis of infectious diseases. The goal on diagnostic microbiology is to offer rapid and accurate detection on specific pathogens within a clinical specimen in an appropriate time period. Genomic sequences representing a specific group of infectious agents or particular strains are often selected on symptom basis of detection task for synthesizing or cloning as probes to detect the presence of specific agents through hybridization. Unique probes can assure the accurate identification of a specific pathogen. However, the rather small amount of DNA or RNA within the clinical specimen may require PCR amplification techniques to improve the sensitivity of pathogen detection.

Conventional PCR primers are in unique sequences, therefore one pair of primers can only selectively amplify one target gene. In the practical detection case of pathogen groups, amplifications of multiple target genes from a set of infectious agents are necessary. In addition, clinical samples usually contain limited genetic materials from various organisms (including host cells). A large number of primers are required in conventional PCR to amplify entire sympton-related pathogens (SRP) target sequences specifically and sensitively that, however, installs high synthesis cost and practical limitation over differential detection. It has been proposed that, with carefully designed multiple-use primers (mu-primers), the number of primers can be reduced significantly (Fernandes and Skiena, 2002). The concept of mu-primers can also be extended to work across with distantly related organisms.

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

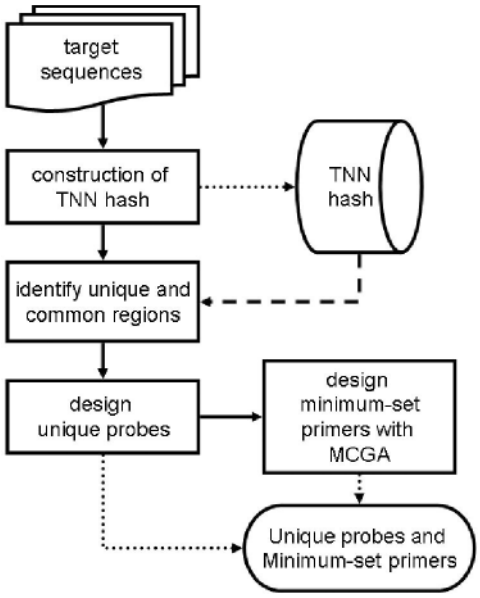


**Fig. 1.** Overall flow of our integrated primer/probe design algorithm.

Selected gene sequences from several pathogen genomes can be amplified with mu-primers (Fernandes and Skiena, 2002) at a greatly reduced number of overall primers required. A large number of SRP target sequences from infectious pathogens may be amplified and detected simultaneously on physical presence in clinical specimens with integrated platform of minimum-set mu-primers amplification (MMA) and probes-array hybridization (PAH). The mu-primers thus enhance the sensitivity of diagnostic PCR, as well as the unique probes enhance the specificity. Current approaches use heuristic algorithms to reduce the number of primers (Fernandes and Skiena, 2002), but their results are either time-consuming or unsatisfactory in solution quality.

In this paper, we propose a new algorithm to design minimum-set primers and unique probes. The flow of our algorithm is shown in Figure 1. We have aimed to accomplish optimum reduction of mu-primers within a reasonable amount of time. By our carefully concerted algorithms, common and unique sections among a set of target sequences are identified. We arbitrarily extend the initial nucleation event of primer annealing reaction with a tetra-nucleotide nucleation (TNN)-hash. Combined with scoring criteria, TNN-hash serves as an index for fast identification of common and unique sections on target sequences for the design of mu-primers and unique probes, respectively. We treat the minimum-set mu-primer design problem as a set covering problem (SCP), and use modified compact genetic algorithm (MCGA) to solve it. Compared with other approaches, our algorithm significantly reduces the number of designed mu-primers required to amplify a large number of target sequences at a significantly shorter period of time. Computationally, the efficiency of our algorithm has been verified with simulation trials on 12 669 sequences. Practically, our algorithm with melting-temperature ($T_m$)-equalization among minimum-set mu-primers by artificial linker sequences has also been applied in the differential detection on nine plant viruses with success for our integrated MMA-PAH differential detection system.

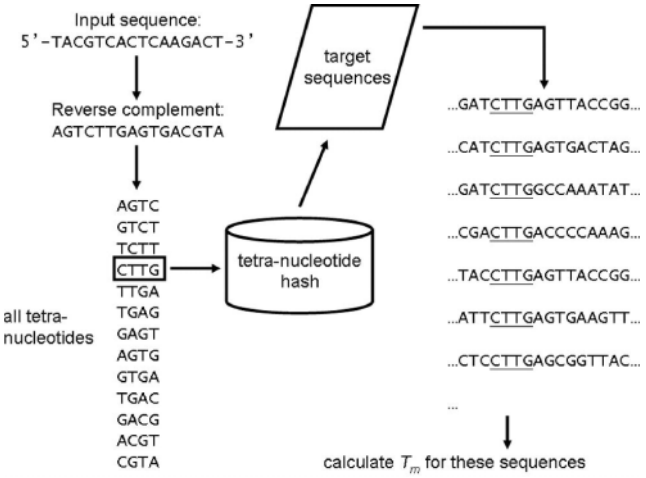


**Fig. 2.** TNN-hash for fast filtering on primer/probe candidates.

## 2 DESIGNING MULTIPLE-USE PRIMERS AND UNIQUE PROBES

Despite that the purpose of degenerate primers and specific probes are different, common techniques are applied to design the mu-primers and unique probes based on their common annealing characteristics with target sequences. In general, the melting temperature ($T_m$) of primers/probes should be calculated (Lockhart *et al*., 1996) and the secondary structures of oligonucleotide must be considered for avoiding dimer and hairpin loops (Wang and Seed, 2003). Most importantly, our algorithm further extends the DNA annealing theory of initial exact-match nucleation (Wetmur and Davidson, 1968) and installs arbitrary TNN hash as a first-pass filter. The regions on target sequences with the most or least shared TNN index patterns are identified as common or unique regions for designing mu-primers or unique probes, accordingly. To reduce the complexity of mu-primer design, we restrict the practical length of designed mu-primers to 12 nt. Linker segments (8mer in length) are designed to respective 12mer mu-primers afterward in order to converge resulted 20mer mu-primers within a small $T_m$ range around 55°C.

### 2.1 Tetra-nucleotide nucleation

For entire target sequences, we build arbitrary TNN-hash for efficient identification of common and unique sections. The hash contains the occurrences of each possible TNN and their respective locations on the target sequences. Based on the initial exact-match nucleation theory for annealing primer/probe onto target sequences, we have assumed that common sections of target sequences contain TNN index patterns with the most occurrence, and unique regions with the least counts.

With each input target sequence, both forward and reverse-complement sequences are generated and looked up in the TNN-hash table. For each TNN, the $T_m$ around each occurrence of the sequence is calculated. These $T_m$ values are used to fast filter sequences, which will (will not) be annealed to the input sequence as illustrated in Figure 2.

With TNN-hash applied throughout the design process, the candidate sequences either for mu-primers-to-linkers or for probes are rapidly screened and evaluated. A set of candidate sequences

are searched based on entries in the TNN-hash, and the resulted set is further screened with their $T_m$ against the original query sequences.

### 2.2 Selecting common and unique sections

Among the shared aspects of primer/probe designs, identifying common and unique regions on the target sequences is one of the major tasks. Several criteria are applied for the selection of common and unique sequences on the target sequences. We have inevitably observed that common sections and unique sequences are likely interspersed. That is, a unique region in a sequence is likely surrounded by common regions which are possibly present among several other sequences of target dataset. On this assumption, our algorithm selects common and unique sections with TNN index of high or low occurrences located on nearby regions of any given target sequences as to be mu-primer and unique probe candidates.

We have used two values to indicate the density of unique TNN in a region ($U_d$) and aggressions of these unique TNN ($U_a$) (Chang and Peck, 2003) as shown in the following equations:

$$U_d = \frac{N_{uniq}}{L} \tag{1}$$

$$U_a = \frac{1}{L}\sqrt{\sum_{i=0}^{n} \frac{((x_{i+1} - x_i) - (L/(N_{uniq}+1))^2}{N_{uniq}+1}}, \tag{2}$$

where $N_{uniq}$ is the number of unique TNN in the sequence and $L$ is the length of the region. The position of the initial base of the $i$-th unique TNN is given by $x_i$. Also, $x_0 = 0$, $x_{n+1} = L$, and $n = N_{uniq}$. The high density of unique TNN ($U_d$) and low aggression of unique TNN ($U_a$) have been shown as good indications for designing high specificity probes (Chang and Peck, 2003). Regions with large $U_d$ values are selected; these regions are much more specific to the given sequence. If the distribution of unique TNN is uniform in the sequence, the value of $U_a$ will approach 0. A lower $U_a$ value indicates that unique TNNs overlap less with each other. Therefore, we propose a modified score for the selection of common and unique regions on target sequences as follows:

$$S = \alpha U_d + \beta \frac{1}{U_a}. \tag{3}$$

The two weights $\alpha$ and $\beta$ can be arbitrarily assigned for selection of common and unique regions, respectively. In this work, we chose equal weights for the two values. If the common regions are selected, the score ($S$) should be minimized. Likewise, the selected region is more unique if the score is higher.

### 2.3 Thermodynamic parameters estimated with nearest-neighbor model

Thermodynamic parameters for hybridizing mu-primers or unique probes to target sequences are estimated with the nearest-neighbor model. The thermodynamic parameters include melting temperatures ($T_m$), free energy ($\Delta G$), enthalpy ($\Delta H$), and entropy ($\Delta S$). The nearest-neighbor model has been validated on accurately estimating thermodynamic parameters (Rahmann and Grafe, 2004; Tanaka *et al*., 2004), which are calculated from hybridization of consecutive di-nucleotides. For example, the enthalpy is calculated as follows:

$$\Delta H = \Delta H_{init} + \sum (\Delta H_{nn}), \tag{4}$$

where $\Delta H_{nn}$ is the enthalpy of hybridization between two di-nucleotides and $\Delta H_{init}$ is the enthalpy for initiation of a DNA duplex. Other thermodynamic parameters (free energy and entropy) are recursively estimated in similar ways. The $T_m$ of annealed sequences are calculated as follows (Sugimoto *et al*., 1996):

$$T_m = \frac{\Delta H}{\Delta S + R \ln (C_T/\alpha)}, \tag{5}$$

where $\Delta H$ and $\Delta S$ are enthalpy and entropy estimated with the nearest-neighbor model, $R$ is the gas constant, $C_T$ is the molar concentration of the oligonucleotide and $\alpha$ equals 4. Mismatches are also considered in the calculation of $T_m$.

## 3 OPTIMIZING MINIMUM-SET MU-PRIMERS WITH SCP AND MCGA

The reduction of mu-primers will significantly enhance the feasibility of multiplex PCR and reduce the cost of diagnostic PCR or microarray analysis. Here we formulate the mu-primers reduction as a constrained SCP. We ought to optimize the number of mu-primers so that we can apply standard MMA for all the task-related target sequences. We use MCGA to solve SCP. Genetic algorithms have been used for designing primers (Wu *et al*., 2004), whereas in this work, our MCGA is for minimizing total mu-primers number with the redundant mu-primers replaced efficiently.

### 3.1 Set covering problem formulated for minimum-set mu-primer design

In conventional PCR experiments, each amplified target sequence requires one pair of primers, one for forward strand amplification and the other for reverse. If $n$ sequences are to be amplified simultaneously, $2n$ primers are required. The required number of mu-primers can be greatly reduced by serving as forward and/or reverse primers among several target sequences. We transform the intended minimum-set mu-primer design problem as a constrained SCP for covering entire target sequences. The constraints are defined so that for each amplified target sequence there must be at least one specific-pair of mu-primers.

The formulation of constrained SCP is as follows:

$$\text{Minimize} \sum_{j=1}^{l} X_j \tag{6}$$

subject to $\forall k$, $1 \le k \le n$, $\exists i \in pair_k$, such that

$$\sum_{j=1}^{l} a_{ij} X_j = \sum_{j=1}^{l} a_{ij}, \tag{7}$$

where $n$ is the number of target sequences, $a_{ij}$ is the element of $n \times l$ zero-one matrix, $l$ is the number of mu-primers and $k$ is the index of mu-primer pairs. If mu-primer $j$ is a primer (forward or reverse) for sequence $i$, then $a_{ij} = 1$, otherwise $a_{ij} = 0$. If mu-primer $j$ is in the solution set, then $X_j = 1$, otherwise $X_j = 0$.

SCP has been proven as an NP-complete problem (Garey and Johnson, 1979). Several heuristic approaches have been proposed for solving SCP. Here we used a genetic algorithm to solve the SCP in a reasonable amount of time. In the next section, we will describe the applied genetic algorithm.

### 3.2 Modified compact genetic algorithm

We have exploited a modified version of compact genetic algorithm (MCGA) to represent the population as a probability distribution over the set of solutions (Harik *et al*., 1999). The fitness evaluation in CGA is based on how the individual approximates the solution. Based on edge replacement, we introduce a local search heuristic into CGA, which is inspired by previous heuristic approaches (Fernandes and Skiena, 2002).

The chromosome is represented as a vector for mu-primers, such as

$$C = (u_1, \ldots, u_j, \ldots, u_t). \quad (8)$$

If mu-primer $j$ is selected, $u_j$ will be 1. Otherwise, $u_j$ will be 0. The chromosome can be seen as an mu-primer set which can satisfy the constraints of the SCP. A probability vector $V$ is used to represent the population:

$$V = (v_1, \ldots, v_j, \ldots, v_t). \quad (9)$$

The element $v_j$ is the probability to select mu-primer $j$. At the initial stage, all the probabilities are set to 0.5 for random selection of mu-primers. All chromosomes are generated according to the probability vector $V$. We use uniform crossover to generate two intermediates from two parents. Local search strategy is applied to these intermediates in order to reproduce valid individuals. These new children will compete with their parents and consequently update the probability vector $V$.

The local search mechanism first makes the intermediates become valid solutions by introducing selected mu-primer pairs to cover all sequences. Then each mu-primer pair is replaced by an alternative mu-primer pair over the same target sequence only if the replacement will decrease the number of mu-primers. The local search process repeats until the number of mu-primers cannot be reduced any further.

The competition between two chromosomes results in a chromosome with better fitness score and the other with inferior fitness. The probability vector is updated based on the result of competition. If mu-primer $j$ presents in the chromosome with a better fitness score but not in the inferior one, the probability for selecting this mu-primer will become $v_{j'} = v_j + 1/n$, where $n$ is the size of population; if mu-primer $j$ presents in the inferior chromosome but not in the one with a better fitness score, then $v_{j'} = v_j - 1/n$; and if mu-primer $j$ presents in both chromosomes, then $v_{j'}$ is not modified. This process will repeat $t$ times for all mu-primers.

Combining the local search heuristic and global search of CGA, the concerted MCGA process terminates when the probability vector $V$ converges on solving the constrained SCP of designing minimum-set mu-primers.

### 3.3 Uneven $T_m$ among minimum-set mu-primers equalized with optimal linker design

The $T_m$ of designed mu-primers may vary in an inconvenient range which often causes practical PCR problems of smearing yield on agarose gel. To make sure that the $T_m$ of all designed mu-primers are within an applicable range, we have adopted a linker-primer design for dual-phase PCR approach. In our approach, a linker will be selected for each mu-primer so the $T_m$ of the mu-primers with linkers may stay in a narrow range around 55°C for performing standard dual-phase PCR assay for validating the applicability of these mu-primers with linkers. Our algorithm first designs minimum-set 12mer mu-primers using MCGA. Additional 8mer linkers are added to the 5′-end of designed mu-primers subsequently without any annealing contribution on mu-primers towards original templates during the touch-down PCR phase at lower $T_m$ of 36–30°C for 13 cycles. The sole purpose of these linkers is to equalize the $T_m$ of minimum-set 20mer mu-primers for steady-state PCR phase at higher $T_m$ of 55°C for 23 cycles. The target sequences upstream to 12mer mu-primers annealing sites are filtered with the TNN-hash to exclude linkers similar to both target sequences and designed mu-primers. The constraints on the selection of linkers can be illustrated with the following equation:

$$\forall L \notin (T \cup P), \quad (10)$$

where $L$ is the set of candidate linkers, $T$ is the set of target sequences with potentially contaminated genomic materials and $P$ is the minimum-set 12mer mu-primers. Also

$$\forall PL \notin (T), \quad (11)$$

where $PL$ is the 20mer set of mu-primers with linkers.

**Table 1.** Comparative solution qualities and performances of applicable algorithms on SCPs

| Algorithms | Deviation of solution (%) | Time |
|---|---|---|
| Direct genetic algorithm | 0.02 | 2583.52 |
| Lagrangean-based heuristic | 0.00 | 94.83 |
| Indirect genetic algorithm | 0.05 | 96.69 |
| Linear time heuristic | 4.02 | 15.50 |
| Densest subgraph heuristic | 0.00 | 418.60 |
| MCGA | 0.00 | 44.33 |

## 4 RESULTS

We have tested the capability of MCGA in solving SCP by comparing it with several other algorithms (Table 1). These algorithms have been applied to several standard benchmark SCP problems (Beasley, 1990) for evaluating relative performances and solution qualities. Likewise, we apply MCGA on designing minimum-set mu-primers, and compare our method with two published heuristics, including linear time heuristic (LTH) and densest subgraph heuristics (DSH) (Fernandes and Skiena, 2002). Despite that these two heuristics originate from different purposes, LTH can reduce the mu-primer set in linear time, but the reduction is not as good as DSH. Moreover, the time complexity of DSH is much higher than that of LTH. Our results illustrate that the outperforming MCGA method is much faster than DSH and the reduction rate is higher than LTH (Table 2).

We have consequently performed a simulation of genome-wide amplification for selected organisms, and compared our MCGA-based method with these two heuristics. We then composed our design of minimum-set primers and unique probes for the differential detection on nine plant viruses.

### 4.1 MCGA for set covering problem

We have compared several algorithms for SCP, including direct genetic algorithm (Beasley and Chu, 1996), Lagrangean-based heuristic (Caprara *et al*., 1999), indirect genetic algorithm

**Table 2.** Comparison on primer reduction among two heuristics and MCGA

| Organism | No. of genes[b] | Linear time heuristic[a] | | Densest subgraph heuristic[a] | | MCGA | |
|---|---|---|---|---|---|---|---|
| | | Reduction[c] (%) | Time (s) | Reduction[c] (%) | Time (s) | Reduction[c] (%) | Time (s) |
| *Schistosoma mansoni* | 817 | 20.85 ± 0.17 | 8.03 | 24.92 ± 1.18 | 30 151.11 | 24.36 ± 0.20 | 1824.70 |
| *Medicago truncatula* | 4466 | 38.61 ± 0.29 | 18.52 | 42.22 ± 1.12 | 74 549.02 | 42.72 ± 0.12 | 3121.42 |
| *Hordeum vulgare* | 11 180 | 63.31 ± 0.27 | 1069.92 | 71.80 ± 2.14 | 227 001.40 | 70.57 ± 0.07 | 7294.19 |
| *Ciona intestinalis* | 12 669 | 55.21 ± 1.65 | 1765.03 | 63.98 ± 4.59 | 467 867.35 | 68.00 ± 0.10 | 12 532.50 |

[a]Implemented as described in the reference article (Fernandes and Skiena, 2002).
[b]Number of genes which can be amplified with appropriate primer pairs.
[c]Percent of reduced primers, averaged over 30 repeated runs.

(Aickelin, 2002), linear time heuristic (Fernandes and Skiena, 2002), densest subgraph heuristic (Fernandes and Skiena, 2002) and our MCGA. These listed algorithms have been applied to 60 standard test sets (Beasley, 1990). The sizes of these test sets range from 200 × 1000 to 1000 × 10 000; and their densities (portions of rows covered by columns) range from 2 to 20%. The averaged results over 10 trial runs for each problem are summarized in Table 1. These comparisons are made on a Pentium II 450 MHz PC. Deviation is the differences between the optimum solution and the solution found by respective algorithms. Smaller deviations indicate that the solution qualities are closer to optimum and can be taken as better solutions.

From the results shown in Table 1, it is clear that MCGA is the one with both best performance and best solution qualities. Linear time heuristic has best performance, but the deviation is too large compared with the solutions of other algorithms. With MCGA proven as better SCP solution, we applied MCGA to design minimum-set mu-primers with superior result.

## 4.2 Simulations

For large-scale analysis such as genome-wide microarray assay, the number of primers and probes are usually the limiting factor, since that the $T_m$ variations of primers and the cost of the microarray experiments increases with the number of target sequences. Our MCGA-based method can reduce the number of mu-primers required to amplify the entire set of target sequences. We have tested four organisms of different genome sizes. Genome sequences for *Schistosoma mansoni*, *Medicago truncatula*, *Hordeum vulgare* and *Ciona intestinalis* are retrieved from NCBI UniGene database (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=unigene).

The comparison among LTH, DSH and MCGA is made in a Pentium 4 2.6 GHz PC running Linux operating system. The $T_m$ range for the genome-wide PCR primers is 37–43°C. Each of the three methods has been repeated for over 30 times. The averaged results of 30 runs including reduction rates, standard deviations and average time used are summarized in Table 2.

From Table 2, we can see that LTH is the fastest algorithm among the three. MCGA is faster than DSH but slower than LTH over an order of magnitude. The mu-primer reduction rates of DSH and MCGA are comparable, both much better than that of LTH. In the case of *C.intestinalis*, DSH shows 63.98% reduction of mu-primers, whereas MCGA reduces 68% of mu-primers required to amplify 12 669 sequences. MCGA uses only 2.68% of the time used by DSH algorithm. That is, comparing with DSH, MCGA may save more than five days when applied on *C.intestinalis*.

**Table 3.** Related pathogen targets of nine plant viruses from four genera

| Virus | Length | Genus | Description |
|---|---|---|---|
| DsMV | 1826 | *Potyvirus* | Dasheen mosaic virus |
| TuMA | 1757 | *Potyvirus* | Turnip mosaic virus |
| ZaMMV | 1845 | *Potyvirus* | Zantedeschia mild mosaic virus |
| ZaMV | 1600 | *Potyvirus* | Zantedeschia mosaic virus |
| PVA | 1682 | *Potyvirus* | Potato virus A |
| CNV | 4701 | *Tombusvirus* | Cucumber necrosis virus |
| TBSV | 4776 | *Tombusvirus* | Tomato bushy stunt virus |
| HCRSV | 3910 | *Carmovirus* | Hibiscus chlorotic ringspot virus |
| BaMV | 6366 | *Potexvirus* | Bamboo mosaic virus |

From these results, we can conclude that MCGA exerts a good balance on achieving both performance and solution quality. The solution quality of MCGA is much better than that of LTH. With better performance and comparable reduction rate with that of DSH, the MCGA is an excellent alternative to the exemplified heuristics for designing minimum-set mu-primers.

## 4.3 Differential detection on nine plant viruses

Our minimum-set mu-primer design algorithm has shown superior results with simulation on genome-wide PCR primer design. In this section, we will combine our minimum-set mu-primer and unique probe design algorithm for practical differential detection on nine plant viruses from four genera as listed in Table 3. Our designed primers and probes are verified in wet laboratory with both MMA based on standard dual-phase PCR assay and PAH. In conventional PCR, 18 primers would be necessary for complete amplification of 9 target sequences. Our designed minimum-set of 11 mu-primers (Table 4) from 11 067 candidates for complete amplification of 9 target sequences reveals 39% reduction rate of mu-primer number. The sequences of 11 mu-primers with linker and their respective $T_m$ for dual-phase PCR assay are provided.

Except for LMu-01, LMu-02, LMu-05, and LMu-08 mu-primers, the rest of the mu-primers in Table 5 are shared by more than one target sequences. For example, LMu-04 mu-primer is the forward primer for Bamboo mosaic virus (BaMV) and also acts as the reverse primer for Hibiscus chlorotic ringspot virus (HCRSV).

We have performed wet laboratory experiments to verify our designed mu-primers and unique probes. The first verification is on whether the designed mu-primers work specifically with

**Table 4.** Minimum-set of 20mer LMu-primers with $T_m$-equalization linker (L, 8mer) onto mu-primer (Mu, 12mer) for differential detection against 9 target viruses with standard dual-phase PCR assay

| LMu primer | 20mer sequence (5′-linker-Mu-3′) | Mu. $T_m$ (°C)[a] | Lmu. $T_m$ (°C)[b] |
|---|---|---|---|
| 1 | CAGTAGGG-TCGAATTTCCAA | 34.8 | 57.7 |
| 2 | ATATAAGG-GTAGCGAGTGCA | 34.9 | 55.1 |
| 3 | TACTATGG-GCTGCTTTCATC | 33.0 | 55.5 |
| 4 | TATATAGC-GAAAGAGCAGCC | 36.4 | 55.1 |
| 5 | ACTATAGC-CAGCAACAGCAG | 35.4 | 55.9 |
| 6 | TACTAGGG-TGTACGCCTCTG | 34.7 | 55.5 |
| 7 | ATATAGGG-AGAAAGGCAAGG | 36.2 | 55.1 |
| 8 | TACTAGGG-TTGTTGGAATGG | 34.4 | 55.1 |
| 9 | TACTATGG-CCTTAGCATTGG | 34.2 | 55.5 |
| 10 | ATATAGGG-ATGAGGACAGGG | 35.0 | 55.5 |
| 11 | TACTATGG-TCTTGGAGTGGG | 36.7 | 55.1 |

[a]The variable $T_m$ of 12mer mu-primers of minimum-set.
[b]The converged $T_m$ of 20mer mu-primers with $T_m$-equalization linker.

**Table 5.** Integrated primer and probe design for differential detection on nine plant viruses with MMA yields of standard dual-phase PCR and PAH

| Virus | $T_m$ (°C)[a] | Start[b] | Forward[c] | Reverse[d] | Size[e] |
|---|---|---|---|---|---|
| DsMV | 86.7 | 854 | LMu-06 | LMu-07 | 414 |
| TuMV | 90.0 | 857 | LMu-03 | LMu-07 | 592 |
| ZaMMV | 84.3 | 739 | LMu-02 | LMu-05 | 832 |
| ZaMV | 90.0 | 376 | LMu-03 | LMu-11 | 982 |
| PVA | 87.5 | 376 | LMu-03 | LMu-11 | 789 |
| CNV | 84.3 | 290 | LMu-09 | LMu-10 | 856 |
| TBSV | 83.4 | 310 | LMu-09 | LMu-10 | 856 |
| HCRSV | 90.0 | 1465 | LMu-04 | LMu-08 | 1208 |
| BaMV | 92.5 | 1874 | LMu-01 | LMu-04 | 1613 |

[a]The melting temperature of the unique probe.
[b]The start base position of the unique probe.
[c]The forward primer used.
[d]The reverse primer used.
[e]The length of sequence amplified by the specific primers, in bp.

standard dual-phase PCR as described in Section 3.3. The specific-pair mu-primers amplification (SMA) with specifically annealed target sequence yields expected DNA band as shown in TuMV case (Fig. 3). In MMA, the specific-pair of LMu-03 and LMu-07 mu-primers amplifies TuMV target sequence to reveal expected DNA band without any interfere of the co-existing mu-primers through our standard dual-phase PCR assay.

With paired lanes for each viral target sequence, the '2' lane is the SMA yields with specific-pair mu-primer (10 pmol) and the 'All' lane is MMA yields with minimum-set mu-primers (10 pmol). The banding sizes among all of the '2' lanes are consistent with our algorithmic design, whereas the band intensities in the 'All' lanes are often less strong due to less usable mu-primers within minimum-set (10 pmol × 2/11). The identical bands of paired lanes (enclosed in solid-line frames) may verify that our algorithm-designed mu-primers should work efficiently without annealing to non-specific target sequences. However, unexpected sizes of yield bands are present in HCRSV case (enclosed in dashed boxes). With PVA case, the intended band in the 'All' lane is almost invisible on gel (Fig. 3).

Before resolving the indicated cases on PVA and HCRSV, we verify whether the intended target sequences are present in the MMA yields labeled for PAH (Fig. 4) onto the entire probes panel arrayed in nine boxes of triplicate-dots format. Table 5 lists the nine target sequences of 50mer probes with qualitative specificity at high $T_m$, respective mu-primers pair, and amplified sequence length. Regardless of the almost invisible and unexpected bands in PVA and HCRSV cases, our differential detection system of integrated MMA and PAH may specifically and efficiently identify the target sequences among various organisms with full success as revealed in Fig. 4.

With higher and lower target-abundance (20, 2, 0.2, 0.02 fmol) for preliminary sensitivity assay to resolve the PVA and HCRSV cases (data not shown), the results indicate that the sensitivity of designed minimum-set mu-primers could be <0.02 fmol despite the unexpected DNA bands on gel between the threshold target-abundance of 2 and 0.2 fmol (Fig. 5A). At the higher target-abundance of 2.0 fmol, the MMA yields of PVA and HCRSV by standard dual-phase PCR has revealed almost identical bands as expected which may complete the specificity validation on designed minimum-set mu-primers. However, at the lower target-abundance of 0.2 fmol which might likely be similar to the situation of clinical specimen, the MMA yields of PVA and HCRSV have again revealed unexpected band patterns (Fig. 5A) as shown in Figure 3. Southern-blot hybridization (SBH) data with 50mer HCRSV probe (Fig. 5B) have clearly enlightened almost identical band patterns with minor intensity variation due to different mu-primer availability regardless of higher and lower target-abundance, which may serve as strong specificity validation on designed minimum-set mu-primers while in comparison with the unexpected band patterns of SMA and MMA yields on gel (Fig. 5A) at lower target-abundance.

Rather, the unexpected sizes of shorter and longer yields at lower target-abundance (Fig. 5) may be due to the forced priming opportunistically driven by the imbalanced kinetics of PCR reaction with over-saturated primers and polymerase and relatively low target-abundance for the shorter yields and due to the promiscuously concatenated DNA synthesis of large and small ones for the longer yields. While proving that unexpected sizes of yield bands impose little or no effect on sensitivity and specificity of differential detection as shown in Figures 4 and 5, this paper shall reasonably establish the successful validation status for our integrated differential detection system of MMA and PAH with mu-primers and unique probes optimally designed by our PDA-MS/UniQ system.

## 5 CONCLUSIONS AND DISCUSSIONS

In this work, we have developed an integrated primer/probe design algorithm with the score-kernel of TNN-hash for differential detection on assorted pathogens simultaneously. We have formulated the minimum-set mu-primer design problem as a SCP which is solved by MCGA with successful reduction rate on the mu-primers number required for complete amplification on assorted sequences (Fig. 1). From the results of computational simulation, we have validated that our MCGA-based method is faster than the best DSH while with comparatively better reduction rates (Tables 1–2). With optimally designed mu-primers and unique probes by our integrated algorithm (Table 5), we have practiced our differential detection

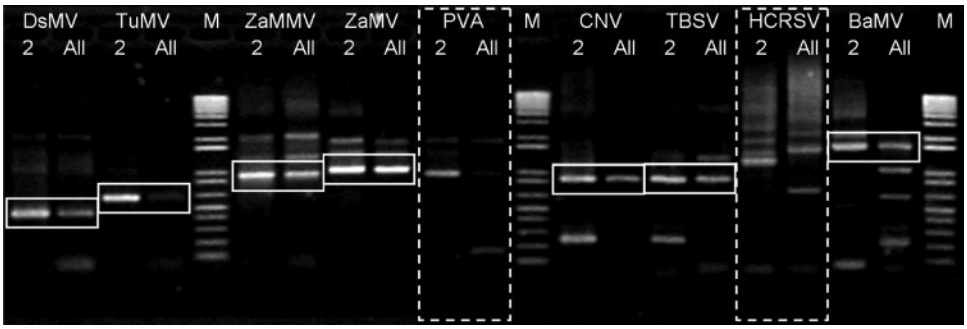


**Fig. 3.** Specificity of SMA and MMA (10 pmol total in '2' and 'All' lanes) verified with respective viral targets (~0.5 fmol) by standard PCR assay and gel electrophoresis. Except for PVA and HCRSV in dashed boxes, identical sizes between SMA and MMA yields of amplified viral targets are shown on 1.5% agarose gel after standard dual-phase PCR assay of the touch-down phase (36–30°C × 13 cycles, −0.5°C each) and the steady-state phase (55°C × 23 cycles). 'M' lanes are 1 kb plus DNA ladders (Invitrogen, Carlsbad, CA, USA).

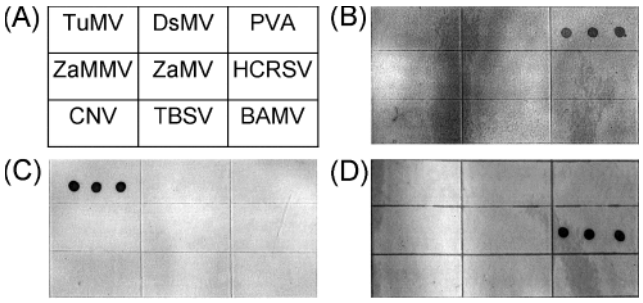


**Fig. 4.** Specificity of differential detection on nine viral targets verified with MMA yields of respective viral targets with DIG-labeling (Roche, Mannheim, Germany) and 45°C PAH of complete probes panel in triplicate-dot format. PAH layout for the differential detection panel of nine viral targets is illustrated in (**A**). For brevity, only three MMA yields of (**B**) PVA, (**C**) TuMV and (**D**) HCRSV targets are applied for verifying the efficient differential detection system of MMA-PAH on the amplified viral targets in MMA yields of standard dual-phase PCR assays, respectively. Hybridized MMA yields with DIG-labeling are detected and indicated with anti-DIG McAb-AP and NBT/BCIP color kit (NEB, Beverly, MA, USA).

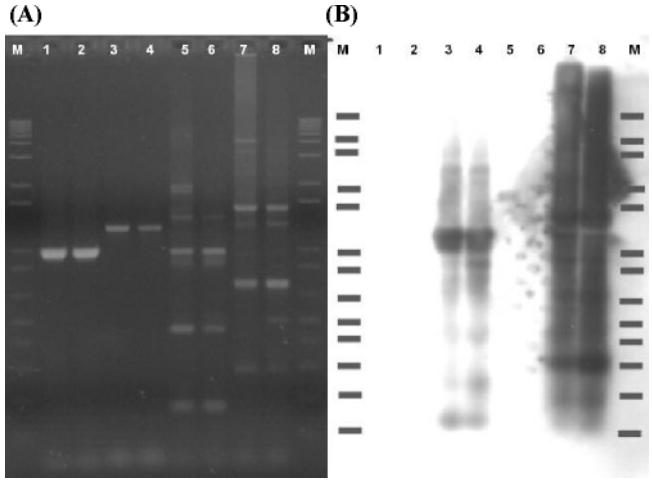


**Fig. 5.** Sensitivity and specificity of integrated MMA-PAH differential detection system verified on **(A)** target dose and **(B)** SBH with the SMA and MMA yields of PVA and HCRSV targets. In **(A)** for verifying sensitivity on target doses, PVA (2.0 and 0.2 fmol, in lanes 1–2 and 5–6) and HCRSV (1.0 and 0.1 fmol, in lanes 3–4 and 7–8) are applied for SMA (lanes 1,3,5 and 7) and MMA (lanes 2,4,6 and 8) yields of standard dual-phase PCR assays (10 pmol primers total) analyzed on 1.5% agarose gel with 'M' lanes of 1 kb plus DNA ladders. In **(B)** for verifying specificity on SMA and MMA yields, SBH is hybridized at 45°C with designed 50mer HCRSV probe of 3′-end DIG-labeling (Roche). The DIG-labeled probes hybridized upon SBH is detected and signal developed with anti-DIG McAb-AP and CSPD/CDP-star chemiluescent kit (NEB) as shown with schematic 'M' lanes of 1 kb plus DNA ladders.

system of integrated MMA and PAH on nine plant viruses with full success (Figs 3–5). Based on the results, our differential detection system of integrated MMA-PAH may identify the presence of target sequences within specimens by MMA for amplifying both true-positive bands and false-positive bands on gel and by PAH for hybridizing true-positive bands out of the mixture with false-positive bands.

For large-scale primer reduction simulated with LTH and DSH, our MCGA outperforms LTH in terms of reduction rate and solution quality. Likewise, while achieving comparable solution quality to DSH, our MCGA may perform with much smaller time complexity despite that the reduction rate of MCGA is slightly lower than DSH in certain cases (Table 2). However, the standard deviation of 30 repeated runs with MCGA is much smaller than those with DSH and LTH (Table 1). The result implies that the MCGA-based method is more robust and capable of achieving consistent reduction rates. In real-world applications, a more robust algorithm is more practical and more applicable towards optimizing differential detection system. In short, the specificity and sensitivity of standard dual-phase MMA and PAH has been successfully and completely validated with optimally designed unique probes and minimum-set mu-primers at either higher or lower target-abundance (Figs 3–5). It is noteworthy that the $T_m$-equalization linkers of respective mu-primers has potentiated the feasibility of standard dual-phase PCR assay at a practical $T_m$ steady-state for producing remarkably specific SMA and MMA yield bands on gel in contrast to the possible smearing results on gel.

The MMA yields of standard dual-phase PCR on nine plant viruses (Table 5) may reveal almost invisible and unexpected results in PVA and HCRSV (Fig. 3). Regardless, our PAH in triplicate dots has nonetheless specifically detected the evident presence of

amplified target sequences despite the almost invisible and unexpected bands in both cases of PVA and HCRSV (Fig. 4). Notably, our integrated MMA-PAH system can efficiently identify the expected true-positive dots by PAH (while identifying true-positive band and rescuing false-negative bands on gel and on blot), despite that the MMA yields is mixed with unexpected false-positive bands and unknown false-negative bands on gel as shown in Figure 5. The minimum-set mu-primers may optimize the standard dual-phase PCR with perfect multiplex feasibility for revealing true-positive false-positive and false-negative MMA yield bands on gel. In addition, the mu-primers of PVA (LMu-03 and LMu-11) are shared with ZaMV (LMu-03 and LMu-11) and TuMV (LMu-03), whereas in both cases the shared mu-primers successfully reveal specific SMA and MMA yield bands in accord with algorithm design. With mu-primers of HCRSV (LMu-04 and LMu-08), the shared mu-primer of BaMV (LMu-04) also reveals intended bands successfully (Fig. 3). Moreover, the unexpected bands on gel commonly seen at lower target-abundance of PVA and HCRSV are verified with SBH (Fig. 5B) to show little or no effect over the HCRSV probe-specific bands in MMA yields (true-positive bands on SBH) despite the unexpected sizes of shorter and longer bands (rescuing false-negative bands along with excluding false-positive ones on gel by SBH). The integrated differential detection system of MMA-PAH with algorithm-designed primers/probes is successfully applicable towards differential detection on nine plant viruses, whereas the formation mechanism of unexpected MMA yields herein shall be resolved with future efforts into sequencing the unexpected bands. In addition, quantitative function of our MMA-PAH differential detection system may be satisfactorily installed in part with future efforts of using introduced control sequences at specified target-abundance for paralleled linear amplification.

With regard to the high false-negative detection rate of conventional PCR on RNA viruses, it is not appropriate to speculate that in conventional PCR cases with unique primers the unexpected bands at lower target-abundance would be as commonly seen as in our system. Regardless, any unexpected bands on gel in conventional PCR cases shall be falsely translated as negative result without applying PAH similar to our integrated system for accurate detection. On the contrary, our standard dual-phase PCR of MMA-PAH can efficiently identify true-positive bands out of the MMA mixture with unexpected false-positive bands and unknown false-negative bands (Figs 4 and 5). Mu-primers at probably conserved common sequences may serendipitously decrease the high false-negative detection rate caused by the unique primers designed at highly mutated segments of RNA virus genome for conventional PCR detection.

The capability to simultaneously detect various pathogens with integrated MMA-PAH platform could be valuable in the biomedical industry and molecular biology studies. Nevertheless, this method may be applied to develop diagnostic chips of automated rapid tests for differential detection on task-related genes. Evidently, our integrated PDA-MS/UniQ system in part may practically design minimum-set mu-primers required to amplify large-scale target sequences, respectively, with specific-pair mu-primers for each target sequence preparation reaction at higher target-abundance, thus saving capital investment. For expensive genome-wide and comparative genomic analysis with microarray, our integrated PDA-MS/UniQ system may thus ameliorate one of the major limitation factors on the stacking cost of expensive oligonucleotide synthesis from the potential microarray applications in various research topics despite the decreasing price of DNA synthesis. Alternatively, advanced microarray or macroarray analysis on the identified target genes set of statistic inference especially at lower target-abundance can be simultaneously and quantitatively analyzed by our MMA-PAH platform upon linear amplification for expression ratio conservation on each target genes with both internal control genes and putative minimum-set mu-primers optimally designed by our integrated PDA-MS/UniQ system in conjunction with target archive technologies of cDNA synthesis and RNA amplification.

## ACKNOWLEDGEMENTS

The authors would like to thank the National Science Council (NSC), the Council of Agriculture (COA), and the Department of Industrial Technology (DoIT) of the Ministry of Economic Affairs (MOEA), Taiwan for financial support: NSC 92-2622-E-002-026, NSC 92-2745-B-034-001, COA 93-AS-3.1.2-AD-U1 and MOEA 93-17-A-19-S1-0016. Funding to pay the Open Access publication charges for this article was provided by Institute for Information Industry (III), Taiwan.

*Conflict of Interest:* none declared.

## REFERENCES

Aickelin,U. (2002) An indirect genetic algorithm for set covering problem. *J. Operat. Res. Soc.*, **50**, 1118–1126.

Beasley,J. (1990) OR-library: distributing test problems by electronic mail. *J. Operat. Res. Soc.*, **41**, 1069–1072.

Beasley,J. and Chu,P. (1996) A genetic algorithm for the set covering problem. *Eur. J. Operat. Res.*, **94**, 392–404.

Caprara,A. *et al*. (1999) A heuristic method for the set covering problem. *Operat. Res.*, **47**, 730–743.

Chang,P.-C. and Peck,K. (2003) Design and assessment of a fast algorithm for identifying specific probes for human and mouse genes. *Bioinformatics*, **19**, 1311–1317.

Fernandes,R.J. and Skiena,S.S. (2002) Microarray synthesis through multiple-use PCR primer design. *Bioinformatics*, **18**, S128–S135.

Garey,M.R. and Johnson,D.S. (1979) *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman, San Francisco.

Harik,G.R., Lobo,F.G. and Goldberg,D.E. (1999) The Compact Genetic Algorithm. *IEEE Trans. Evol. Comput.*, **3**, 287–297.

Lockhart,D.J. *et al*. (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat. Biotechnol.*, **14**, 1675–1680.

Rahmann,S. and Grafe,C. (2004) Mean and variance of the Gibbs free energy of oligonucleotides in the nearest neighbor model under varying conditions. *Bioinformatics*, **20**, 2928–2933.

Sugimoto,N. *et al*. (1996) Improved thermodynamic parameters and helix initiation factor to predict stability of DNA duplexes. *Nucleic Acids Res.*, **24**, 4501–4505.

Tanaka,F. *et al*. (2004) Thermodynamic parameters based on a nearest-neighbor model for DNA sequences with a single-bulge loop. *Biochemistry*, **43**, 7143–7150.

Wang,X. and Seed,B. (2003) Selection of oligonucleotide probes for protein coding sequences. *Bioinformatics*, **19**, 796–802.

Wetmur,J.G. and Davidson,N. (1968) Kinetics of renaturation of DNA. *J. Mol. Biol.*, **31**, 349–370.

Wu,J.-S. *et al*. (2004) Primer design using genetic algorithm. *Bioinformatics*, **20**, 1710–1717.