EvoOligo: Oligonucleotide Probe Design With Multiobjective Evolutionary Algorithms

Soo-Yong Shin, In-Hee Lee, Young-Min Cho, Kyung-Ae Yang, and Byoung-Tak Zhang

***Abstract*—Probe design is one of the most important tasks in successful deoxyribonucleic acid microarray experiments. We propose a multiobjective evolutionary optimization method for oligonucleotide probe design based on the multiobjective nature of the probe design problem. The proposed multiobjective evolutionary approach has several distinguished features, compared with previous methods. First, the evolutionary approach can find better probe sets than existing simple filtering methods with fixed threshold values. Second, the multiobjective approach can easily incorporate the user's custom criteria or change the existing criteria. Third, our approach tries to optimize the combination of probes for the given set of genes, in contrast to other tools that independently search each gene for qualifying probes. Lastly, the multiobjective optimization method provides various sets of probe combinations, among which the user can choose, depending on the target application. The proposed method is implemented as a platform called EvoOligo and is available for service on the web. We test the performance of EvoOligo by designing probe sets for 19 types of Human Papillomavirus and 52 genes in the Arabidopsis Calmodulin multigene family. The design results from EvoOligo are proven to be superior to those from well-known existing probe design tools, such as OligoArray and OligoWiz.**

***Index Terms*—Evolutionary multiobjective optimization, EvoOligo, microarray probe design, ϵ-multiobjective evolutionary algorithm (ϵ-MOEA).**

I. Introduction

Deoxyribonucleic acid (DNA) microarrays have widely been used in diverse research and commercial areas where the expression levels of a large number of genes need to be simultaneously monitored. There are currently two basic types of DNA microarray: cDNA microarray and oligonucleotide microarray. cDNA microarray consists of cDNA fragments that are usually longer than 200 nt and microspotted on a solid surface. These cDNA fragments are generated by reverse transcription from mRNAs. In contrast, the oligonucleotide microarray uses synthetic oligonucleotides called oligonucleotide probes, which are relatively short (up to 60 nt) DNA complementaries to the genes of interest. Thus, one can customize the probe sequences on oligonucleotide microarrays. The expression level of a gene is estimated from the florescent signal intensity, which represents the amount of probes that bind (hybridized) to that gene. Since a fixed amount of each probe is spotted on the microarray, the amount of probes hybridized to a gene is proportional to the expression level of that gene. However, if a chosen probe is not specific to its target throughout the whole genome, it may hybridize to wrong genes (cross hybridization), which causes misleading signals. Therefore, the reliability of the information provided by an oligonucleotide microarray depends on the quality of the chosen probe set.

A large number of oligonucleotide probe design methods, as well as criteria, have been suggested in the literature. As for the criteria, most studies have used binding free energy or BLAST match score for nontarget sequences, the melting temperature range, and the secondary structure of probes as the main criteria to choose the optimal probe set [1]–[3]. In addition to these popular criteria, frequency-based selection of probes for cDNA clones [4], information theoretical measure based on Shannon entropy [5], and sequence feature information [6], [7] have been suggested. In contrast to the diversity of the probe evaluation criteria, most of microarray probe design tools share similar probe search methods, which can be categorized as filtering methods. The candidate probes are first generated by scanning through the target sequences. Next, the probes are evaluated according to the specified criteria, and those that do not meet the required threshold values are discarded. Finally, the remaining probes are sorted according to their scores, and the best one is recommended. In this way, each sequence in the set of targets is independently searched one by one.

However, the filtering-by-threshold approach has some drawbacks. First, since the result from this method is usually a set of candidate probes where each target has multiple candidates, the combination of candidates should be determined. Although

Manuscript received May 28, 2008; revised December 11, 2008 and April 14, 2009. First published June 23, 2009; current version published November 18, 2009. This work was supported in part by the Korean Ministry of Education and Human Resources Development under the BK21-IT Program, by the Korean Ministry of Knowledge Economy through the MEC project, and by the Korea Science and Engineering Foundation (KOSEF) through the National Research Laboratory Program funded by the Ministry of Science and Technology under Grant M104000000349-06J0000-34910. The work of S-Y. Shin was supported by the Korean Government (MOEHRD) under Korea Research Foundation Grant KRF-2006-214-D00140. The work of K.-A. Yang was supported by the Korean Government (MOEHRD) under Korea Research Foundation Grant KRF-2008-532-C00031. This paper was recommended by Associate Editor E. Santos Jr.

S.-Y. Shin is with the Medical Information Center, Seoul National University Hospital, Seoul 110-744, Korea (e-mail: syshin@snuh.org).

I.-H. Lee is with the Biointelligence Laboratory, School of Computer Science and Engineering, Seoul National University, Seoul 151-742, Korea (e-mail: ihlee@bi.snu.ac.kr).

Y.-M. Cho is with the University of California San Diego, La Jolla, CA 92093 USA (e-mail: ymcho@bi.snu.ac.kr).

K.-A. Yang is with the Center for Bioinformation Technology (CBIT), Seoul National University, Seoul 151-742, Korea (e-mail: kayang@bi.snu.ac.kr).

B.-T. Zhang is with the Biointelligence Laboratory and the Center for Bioinformation Technology (CBIT), School of Computer Science and Engineering, Seoul National University, Seoul 151-742, Korea (e-mail: btzhang@bi.snu.ac.kr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSMCB.2009.2023078

the candidates are scored, depending on their specificities to the target, the collection of high-scoring probes for each target may not be the best probe set. Therefore, the combination of probes still needs to be optimized. Second, although the probe design requires multiple criteria, most of the tools have used a single score term for filtering. However, He *et al.* [8] mentioned that a single criterion is not stringent enough to eliminate nonspecific probes and that a set of essential criteria must be considered. This supports the multiobjective nature of the probe design. Third, the threshold values for discriminating nonspecific probes are heuristically chosen. However, the recent analyses on popular probe design criteria such as free energy report that it is difficult to establish a clear threshold value and explain their effects on the hybridization intensity [8]–[10].

From the preceding observations, we propose an evolutionary multiobjective optimization approach for the oligonucleotide probe design. First, the stochastic parallel search of an evolutionary algorithm makes it suitable to optimize a combination of probes among the exponential amount of possible combination of candidate probes. Second, we do not have to explicitly set the threshold values as in the simple filtering method by using an evolutionary optimization approach. Therefore, our method does not suffer from inadequate threshold values. Third, the multiobjective evolutionary approach can handle the multiple objectives required in the probe design without aggregating them into a single term as in the filtering methods, where one should carefully adjust a relative weight of each objective. Moreover, it can provide various probe sets at a time while avoiding repetitive work of adjusting aggregation parameters. Lastly, the flexibility of the multiobjective approach makes it easy to incorporate specific requirements of users including addition or modification of the criteria. Based on our previous research [11], the probe design problem has been formulated as a constrained multiobjective optimization task and tackled by an ϵ-multiobjective evolutionary algorithm. In addition, we developed EvoOligo (http://cbit.snu.ac.kr/EvoOligo), which is a web-interfaced platform to find optimal probe sets.

In the succeeding sections, we explain the suggested probe design method in detail. In Section II, we formulate the probe design as a multiobjective optimization task. Section III introduces the EvoOligo system. Section IV provides the probe design results and analyses for the genomes of 19 selected types of Human Papillomavirus (HPV) and 52 genes of the Arabidopsis Calmodulin multigene family. In Section V, the conclusion will be drawn.

II. Multiobjective Probe Optimization

A. Multiobjective Optimization

The purpose of a multiobjective optimization is to optimize multiple conflicting objectives at the same time. The general form of a multiobjective optimization problem (MOP) is

$$\begin{aligned}&\text{optimize} \quad f_i(\mathbf{X}), \quad i = 1, \ldots, M\\&\text{subject to} \quad g_j(\mathbf{X}) = 0, \quad j = 1, \ldots, N\\&\qquad\qquad h_k(\mathbf{X}) \geq 0, \quad k = 1, \ldots, O\\&\mathbf{X} = (x_1, \ldots, x_n), \quad x_l^{(L)} \leq x_l \leq x_l^{(U)}; \quad l = 1, \ldots, n \end{aligned} \tag{1}$$

where f represents the objective, g is the equality constraint, h is the inequality constraint, M denotes the number of objectives, N is the number of equality constraints, and O is the number of inequality constraints. $x^{(L)}$ and $x^{(U)}$ are the lower and upper bounds of decision variable x, respectively. n is the number of variables.

To compare the different solutions in a MOP and judge the superiority of a solution, a relationship called dominance is usually used [12]. It can be defined as follows: In the case of maximization, a solution $\mathbf{X}$ is said to dominate another solution $\mathbf{Y}$ when

$$\begin{aligned}&\forall i \in \{1, \ldots, M\}, \qquad f_i(\mathbf{X}) \geq f_i(\mathbf{Y})\\&\exists i \in \{1, \cdots, M\}, \qquad f_i(\mathbf{X}) > f_i(\mathbf{Y}).\end{aligned} \tag{2}$$

Since there is no priority between the objectives, one solution is thought to be superior to another one only if it is not worse than the other in all objectives and strictly better than the other in at least one objective. If there is no dominating solution, the solutions are "nondominated" to each other and treated as equally good.

The most ideal solution of a MOP is one that dominates the others. However, it is impossible to find a single dominating solution due to the conflicting relation between objectives. Thus, a practical solution of a MOP is to find a set of mutually nondominated solutions (a nondominated set) that approximates the set of efficient solutions (Pareto set). Each solution in the nondominated set corresponds to a different tradeoff among multiple objectives. At the final step, therefore, a decision maker is necessary to select a recommended solution among the tradeoff solutions.

B. Probe Design as a Multiobjective Optimization Problem

The probe design requires particular conditions for successful experimentation. In general, the conditions could be summarized here.

1) The complementary sequence of a probe should appear in its target gene only.
2) The possible nonspecific hybridization (cross hybridization) should be minimized.
3) The secondary structures of a probe such as a hairpin structure should be minimized.
4) The experimental conditions such as melting temperature (Tm) should be as similar as possible.

Conditions 1 and 2 concern the specificity of probes. The first condition defines a basic constraint, which any valid probe set should satisfy. However, even if a probe set satisfies the first condition, cross hybridizations can still be possible. Therefore, the second condition, i.e., minimization of cross hybridization, is necessary. The third condition disapproves any secondary structure of a probe since it can disturb hybridization with its target gene and decrease its sensitivity. The latter is to ensure the efficiency of a probe set, because probes on an oligonucleotide microarray will be exposed to the same experimental conditions.

Having found that the second and third conditions conflict with each other to some extent [11], we formulate the microarray probe design using two fitness functions and one constraint. In a probe design practice, the fourth requirement can be compromised for better specificity and sensitivity. Therefore, it is not considered as one of the objectives but regarded as the final decision criterion to choose the best solution among diverse nondominated solutions from the run of a multiobjective evolutionary algorithm (MOEA).

Before continuing to the formulation of the problem, we introduce the basic notations. We denote a set of n probes by $P = \{p_1, p_2, \ldots, p_n\}$, where $p_i \in \{A, C, G, T\}^l$ for $i = 1, 2, \ldots, n$, and l is the length of each probe. We also denote a set of target sequences by $T = \{t_1, \ldots, t_n\}$, where the target of p_i is t_i for $i = 1, 2, \ldots, n$. We assume that each p_i is complementary to the substring of t_i. A complement of DNA sequence s is denoted as $\bar{s}$. Using these notations, the constraint is formulated as follows:

$$g(P) = \sum_i \sum_{j \neq i} subseq(\bar{p_i}, t_j)$$

$$subseq(\bar{p_i}, t_j) = \begin{cases} 1, & \text{if } \bar{p_i} \text{ occurs in } t_j \text{ at least once} \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

This constraint should be zero from the definition of probe. Two objective functions could be represented as

$$f_{\text{Cross}-\text{Hyb}}(P) = \sum_i \min_{j \neq i} FEnergy(p_i, t_j) \tag{4}$$

$$f_{\text{Self}}(P) = \sum_i FEnergy_{\text{self}}(p_i). \tag{5}$$

$FEnergy(p_i, t_j)$ calculates a minimum free energy required for the most stable hybridization between p_i and t_j. Likewise, $FEnergy_{\text{self}}(p_i)$ calculates a minimum free energy required for the most stable secondary structure of p_i. Therefore, $f_{\text{Cross}-\text{Hyb}}$ and f_{Self} calculate the stabilities of the cross hybridization and the secondary structure of the probe, which disturb the hybridization of a probe with its target gene. Since a smaller free-energy value means a more stable interaction or structure, these functions should be maximized. The values of $FEnergy(\cdot, \cdot)$ and $FEnergy_{\text{self}}(\cdot)$ are calculated using the Mfold program [13], which is based on the nearest neighbor model [14].

From the previous discussion, the probe design problem is formulated as a MOP with two maximization objectives and one equality constraint, i.e.,

$$\begin{aligned} \text{Maximize} \quad & f_{\text{Cross}-\text{Hyb}}(P) \\ & f_{\text{Self}}(P) \\ \text{subject to} \quad & g(P) = 0. \end{aligned} \tag{6}$$

C. ϵ-MOEA for Probe Optimization

Based on the MOP formulation of probe selection in (6), we applied a MOEA to the task of probe optimization. MOEA has also experimentally been proven to be useful in similar domains, such as DNA sequence optimization for DNA computing [15] and multiplex polymerase chain reaction (PCR) assay design [16].

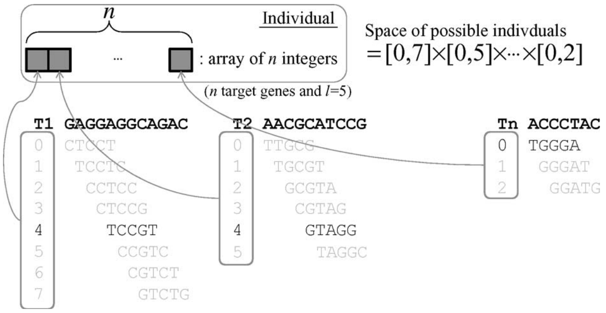


Fig. 1. Individual representation in EvoOligo. The individual consists of n integers representing the starting positions of probes in target genes. On evaluation, corresponding probe sequences are used in (3)–(5).

We adopted the ϵ-multiobjective evolutionary algorithm (ϵ-MOEA), which has shown good performance among diverse MOEAs [17]–[19], as the main optimization algorithm for probe design. It is a steady-state genetic algorithm using ϵ-dominance relation (the most essential feature) and an elite archive to keep the nondominated solutions found so far. The ϵ-dominance is used to maintain a representative subset of nondominated individuals by dividing the whole search space into many grids, whose sizes are defined by ϵ. Only one solution among those from the same grid is kept in the archive. In this way, the minimum distance between the nearest solutions can be guaranteed. The density of the subset can be adjusted by controlling the value of ϵ [17]. In an additive ϵ-dominance relation for the maximization case

$$\begin{aligned} &\mathbf{X} \ \epsilon\text{−dominates } \mathbf{Y} \iff \\ &\qquad \forall i \in \{i, \ldots, M\}, \quad f_i(\mathbf{X}) + \epsilon \geq f_i(\mathbf{Y}) \\ &\qquad \exists i \in \{i, \ldots, M\}, \quad f_i(\mathbf{X}) + \epsilon > f_i(\mathbf{Y}). \end{aligned} \tag{7}$$

Utilizing the ϵ-dominance to select a representative subset of nondominated set and maintaining them in the archive throughout generations, ϵ-MOEA has shown good convergence and diversity performance. We slightly modified the source code provided by Deb (http://www.iitk.ac.in/kangal/codes.shtml) by changing individual representation, fitness functions, and the constraint handling procedure.

Each individual (probe set $\{p_1, \ldots, p_n\}$) is encoded as an array of integers (Fig. 1). The length of the array is n (the number of target genes). The ith element of the array represents the starting point of the ith probe p_i in the ith target gene t_i and has a value between 0 and $|t_i| - l$. $|t_i|$ denotes the length of t_i, and l is the length of each probe. The elements of the array can be different from each other, because the probes can have different starting points in their corresponding target genes. In Fig. 1, there are eight probe candidates for the first target gene (T1), and each is denoted by its starting position in T1. Therefore, the first element of an individual would have an integer value between 0 and 7. Similarly, the second target gene has six probe candidates, and the corresponding element in an individual would have an integer value between 0 and 5. Thus,

```
1) Randomly generate initial set $\mathcal{P}$ of probe sets (population).
2) Sort by domination, and set non-dominated probe sets among $\mathcal{P}$ as archive $\mathcal{A}$.
3) Generate a new probe set $C$ by choosing parents from $\mathcal{P}$ and $\mathcal{A}$.
    a) Choose two probe sets from $\mathcal{P}$ randomly. If any of them violate the constraint, choose the probe
       set with smaller constraint violation. If both satisfy the constraint, choose the dominating probe set.
       Otherwise, choose one at random.
    b) Choose one probe set randomly from $\mathcal{A}$.
    c) Perform crossover and mutation.
4) Evaluate the new probe set $C$ using Eqns. (3) $\sim$ (5).
5) Update archive. Let $\mathcal{R}_1 = \{P|P \in \mathcal{A}, g(P) > g(C)\}$, $\mathcal{R}_2 = \{P|P \in \mathcal{A}, C\ \epsilon\text{-dominates } P\}$, and $\mathcal{R}_3 =$
   $\{P|P \in \mathcal{A}, P\ \epsilon\text{-dominates } C\}$.
    a) If $g(C) \geq 0$ and $\mathcal{R}_1 \neq \emptyset$, let $\mathcal{A} = \mathcal{A} \cup \{C\} - \mathcal{R}_1$.
    b) If $g(C) = 0$, $\mathcal{R}_1 = \emptyset$ and $\mathcal{R}_3 = \emptyset$, let $\mathcal{A} = \mathcal{A} \cup \{C\} - \mathcal{R}_2$. If there are more than one archive
       member in the same grid, keep the dominating probe set. As a tie-breaker, we use the distance to
       the dominating corner of the surrounding grid.
    c) Otherwise, discard $C$.
6) Update the population. Let $\mathcal{S}_1 = \{P|P \in \mathcal{P}, C \text{ dominates } P\}$ and $\mathcal{S}_2 = \{P|P \in \mathcal{P}, P \text{ dominates } C\}$.
    a) If $\mathcal{S}_1 \neq \emptyset$ and $\mathcal{S}_2 = \emptyset$, let $\mathcal{P} = \mathcal{P} \cup \{C\} - \{X\}$, where $X$ is randomly chosen from $\mathcal{S}_1$.
    b) If $\mathcal{S}_1 = \mathcal{S}_2 = \emptyset$, $\mathcal{P} = \mathcal{P} \cup \{C\} - \{X\}$, where $X$ is randomly chosen from $\mathcal{P}$.
    c) Otherwise, discard $C$.
7) Check termination.
    a) If the condition is not met, go to Step 3).
    b) Otherwise, return $\mathcal{A}$.
```

Fig. 2. Pseudocode of ϵ-MOEA for probe optimization.

the total space of possible individuals is the Cartesian product of probe candidate sets for target genes. On evaluation, we temporarily create an $n \times l$ array of integers to represent actual probe sequences. (The probe sequences were more emphasized than others in Fig. 1.) The elements of this temporary array can have a value between 0 and 3, which represents the possible values A, C, G, and T. Then, this temporary array is used to evaluate (3)–(5).

The total procedure of evolutionary probe optimization by ϵ-MOEA is shown in Fig. 2. First, a random population is initially generated (Step 1). Then, we form an archive with the nondominated set (Step 2). Afterward, the algorithm repeats the variation (Step 3), evaluation (Step 4), and maintenance of archive and population (Steps 5 and 6, respectively) until the termination condition is met. In the variation step, parents are selected from archive and population, respectively. We used uniform crossover and point mutation to produce a new probe set. During the maintenance step, keeping feasible individuals is first emphasized since the probe design is formulated as a constrained optimization. The constraint violation is first checked whenever the archive member is updated. An infeasible individual with a smaller violation is preferred over an infeasible one with a larger violation, and a feasible individual is preferred over an infeasible one. The ϵ-dominance check is performed only when both are feasible. In this way, the archive is driven toward the feasible and nondominated region while keeping the representative and proper subset of the nondominant set found so far. When updating the population, only dominance relation is considered to improve the diversity.

III. EvoOligo: Multiobjective Evolutionary Probe Optimizer

The proposed MOEA approach for probe selection is implemented as EvoOligo (http://cbit.snu.ac.kr/EvoOligo/), which is a web-interfaced platform. It is also a part of DNAChipBench (http://cbit.snu.ac.kr/~DNAChipBench/), which is an integrated platform for supporting the whole pipeline related to the design, manufacturing, analysis, and applications of microarrays.

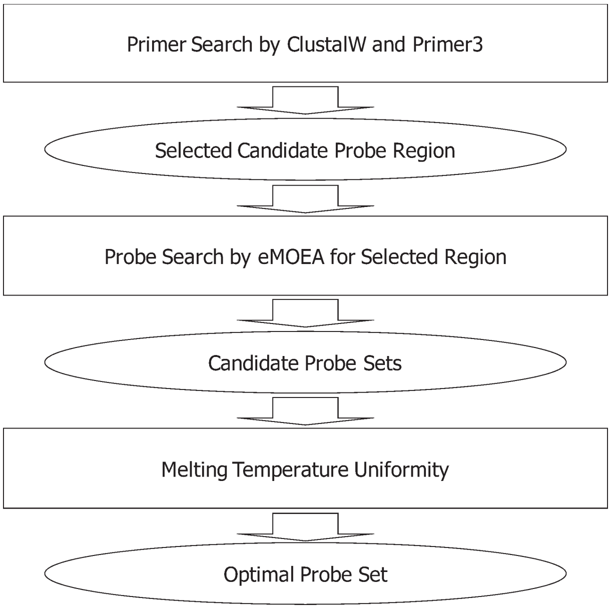
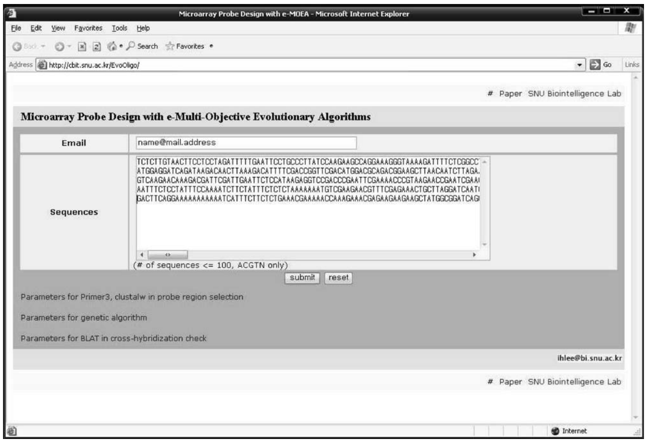

Primer Search by ClustalW and Primer3
Selected Candidate Probe Region
Probe Search by eMOEA for Selected Region
Candidate Probe Sets
Melting Temperature Uniformity
Optimal Probe Set

Fig. 3. Overall steps for probe design in EvoOligo.

The overall EvoOligo procedure is shown in Fig. 3. First, the user should set the parameters and input the sequences via the web interface shown in Fig. 4. The parameters in the web interface are grouped according to their part in the whole procedure, and most of them are directly delivered to the well-known tools used in EvoOligo steps: ClustalW [20], Primer3 [21], Mfold [13], and BLAT [22]. The recommended parameter values proposed by the original authors are available as default. After setting the parameters via the web interface, the users can submit the queries or correct invalid parameters highlighted in red. Then, probe candidate areas are chosen using ClustalW and


Microarray Probe Design with e-MOEA - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Search
Favorites
Address
Go
Links
# Paper SNU Biointelligence Lab
submit
reset
Internet

Address: http://cbit.snu.ac.kr/EvoOligo/

Microarray Probe Design with e-Multi-Objective Evolutionary Algorithms

Email: name@mail.address

Sequences:

```
TCTCTTGTAACTTCCTCCTAGATTTTTGAATTCCTGCCCTTATCCAAGAAGCCAGGAAAGGGTAAAAGATTTTCTCGGCC
ATGGAGGATCAGATAAGACAACTTAAAGACATTTTCGACCGGTTCGACATGGACGCAGACGGAAGCTTAACAATCTTAGA.
GTCAAGAACAAAGACGATTCGATTGAATTCTCCATAAGAGGTCCGACCCGAATTCGAAAACCCGTAAGAACCGAATCGAA
AATTTCTCCTATTTCCAAAATCTTCTATTTCTCTCTAAAAAAATGTCGAAGAACGTTTCGAGAAACTGCTTAGGATCAAT
GACTTCAGGAAAAAAAAAAAATCATTTCTTCTCTGAAACGAAAAACCAAAGAAACGAGAAGAAGAAGCTATGGCGGATCAG
```

(# of sequences <= 100, ACGTN only)

Parameters for Primer3, clustalw in probe region selection

Parameters for genetic algorithm

Parameters for BLAT in cross-hybridization check

ihlee@bi.snu.ac.kr

Fig. 4. Web interface of EvoOligo.

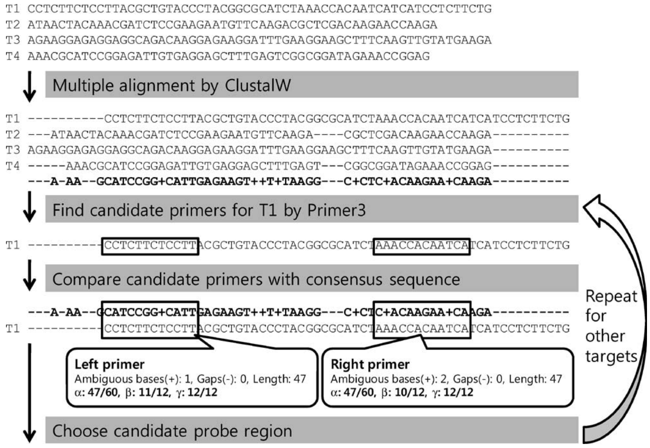
1) Align the entire sequences by ClustalW and extract a consensus sequence. The consensus sequence is represented with the IUPAC ambiguity codes.
2) For every sequence, extract the candidates for the left and right primers respectively by Primer3.
3) Compare the candidate primers from the previous step with the consensus sequence. In this step, the non-degeneracy ratio (β), which exhibits the maximum allowable ratio of ambiguous bases over the primer length, and the non-gap ratio (γ), which reflects the minimum allowable length of gaps, are used as criteria. If the candidates do not satisfy the ratios, they are removed.
4) For every possible pair from Step 3, filter out pairs based on the distance between primers ratio (α), which decides the adjacency of primer pairs to control the size of the probe search space.
5) Choose one from the result of Step 4, which shows minimum variation of melting temperature between the left and right primers.

Fig. 5. Procedure of searching a probe candidate region. The IUPAC ambiguity codes are the standard to resolve the ambiguity of nucleotides in the DNA sequences.

Primer3 to reduce a search space using the procedure shown in Fig. 5. Depending on the purpose of the users, the preprocessing step can be omitted. After preprocessing, ϵ-MOEA searches for a nondominated set of probe sets as in Fig. 2. Finally, one candidate solution is recommended based on melting temperature uniformity. After EvoOligo finishes the optimization, the final nondominated probe sets and the recommended one will be sent to the users by e-mail. A more detailed explanation for each step, except for ϵ-MOEA, is given in the succeeding sections.

A. Preprocessing for the Probe Candidate Region

The preprocessing step is included for efficient probe search since gene sequences are usually very long. In addition, many users are interested in designing oligonucleotide arrays to analyze a specific group of related genes. In such cases, there are usually some conserved regions in the input set of sequences. Therefore, we can remove the conserved regions and concentrate on a subregion between them, which shows wider variation on the sequences.


Multiple alignment by ClustalW
Find candidate primers for T1 by Primer3
Compare candidate primers with consensus sequence
Left primer
Ambiguous bases(+): 1, Gaps(-): 0, Length: 47
α: 47/60, β: 11/12, γ: 12/12
Right primer
Ambiguous bases(+): 2, Gaps(-): 0, Length: 47
α: 47/60, β: 10/12, γ: 12/12
Choose candidate probe region
Repeat for other targets

```
T1 CCTCTTCTCCTTACGCTGTACCCTACGGCGCATCTAAACCACAATCATCATCCTCTTCTG
T2 ATAACTACAAACGATCTCCGAAGAATGTTCAAGACGCTCGACAAGAACCAAGA
T3 AGAAGGAGAGGAGGCAGACAAGGAGAAGGATTTGAAGGAAGCTTTCAAGTTGTATGAAGA
T4 AAACGCATCCGGAGATTGTGAGGAGCTTTGAGTCGGCGGATAGAAACCGGAG

T1 ----------CCTCTTCTCCTTACGCTGTACCCTACGGCGCATCTAAACCACAATCATCATCCTCTTCTG
T2 ---ATAACTACAAACGATCTCCGAAGAATGTTCAAGA----CGCTCGACAAGAACCAAGA----------
T3 AGAAGGAGAGGAGGCAGACAAGGAGAAGGATTTGAAGGAAGCTTTCAAGTTGTATGAAGA----------
T4 -----AAACGCATCCGGAGATTGTGAGGAGCTTTGAGT---CGGCGGATAGAAACCGGAG----------
   ---A-AA--GCATCCGG+CATTGAGAAGT++T+TAAGG---C+CTC+ACAAGAA+CAAGA----------

T1 ----------CCTCTTCTCCTTACGCTGTACCCTACGGCGCATCTAAACCACAATCATCATCCTCTTCTG

   ---A-AA--GCATCCGG+CATTGAGAAGT++T+TAAGG---C+CTC+ACAAGAA+CAAGA----------
T1 ----------CCTCTTCTCCTTACGCTGTACCCTACGGCGCATCTAAACCACAATCATCATCCTCTTCTG
```

Fig. 6. Demonstration of the preprocessing step. In the left primer, α is the length of the probe candidate region (47) divided by the length of T1 gene (60), β is the number of nondegenerate bases [11, since there is only one ambiguous base $(+)$] divided by the length of primer (12), and γ is the nongap bases (12) divided by the length of the primer. In the right primer, only β is different from the left primer since there are two ambiguous bases.

To find the candidate subregion, multiple sequence alignment is combined with the PCR primer search method. Simply, PCR is a procedure to amplify a DNA fragment between two selected areas, and short complementary oligonucleotides that bind to the selected area are called primers. Since a primer serves as a starting point for DNA fragment replication, a primer should not interact with itself or any other primers. In addition, since primers are exposed in the same experimental condition, the chemical properties (mostly the melting temperature) should be as similar as possible. Usually, the region between the selected primers (PCR region) is regarded as a representative substring of the sequence, which could provide good candidates for probe. However, the PCR region might include the conserved regions among gene sequences since the PCR primer selection does not consider other genes in the input set. Therefore, we have to find the PCR region that excludes the conserved region. This goal is achieved by comparing the PCR region with multiple sequence alignment results, which can identify the conserved region.

Fig. 5 summarizes the major steps in searching for a probe candidate region, and Fig. 6 shows an example of the preprocessing procedure. In the first step, we draw out a consensus sequence from the entire input sequences by multiple sequence alignment. The consensus sequence is represented with the International Union of Pure and Applied Chemistry (IUPAC) ambiguity codes, which are the standard to resolve the ambiguity of nucleotides in the DNA sequences. Except for four DNA bases, there are an ambiguous base $(+)$, gap $(-)$, and others in the IUPAC ambiguity codes (http://www.ncbi.nlm.nih.gov/SNP/iupac.html). Then, we compare the primer candidates independently obtained for each target with the consensus sequence and select the appropriate primer pair for each target according to the constraints in Steps 3–5 in Fig. 5. Steps 3 and 4 filter the PCR regions that might contain a conserved region. In detail, Step 3 checks whether the primers are chosen from the conserved region by calculating the ratio of nonambiguous (specified or conserved) bases on the consensus sequence, which corresponds to the primer position over the primer length β and the ratio of bases other than gaps on the consensus sequence over the primer length γ. Step 4 specifies the minimum ratio of distance between the primers over input sequence length α, which controls the size of the probe search space. Step 5 forces to select the primer pair with a similar melting temperature.

In a region where the diversity among sequences is high, most sequences would look different; consequently, most of the consensus sequence will be filled with wild card characters and few nonambiguous bases. Therefore, if the primers are from a highly conserved region (larger β and γ values) and are guaranteed to be separated by some specified distance larger than the distances between the conserved regions (larger α values), the region between them can be specific and exclude the highly conserved regions. The parameters in the preprocessing step should carefully be set, considering the overall similarity of the input sequences. Since the resulting probe candidate regions correspond to the search space of ϵ-MOEA, adjusting these parameters directly affects the search efficiency of EvoOligo. However, there might not be such a specific conserved region for some genes. Therefore, this preprocessing step can be omitted in that case.

B. Selection of Final Probe Set

The advantage of the multiobjective approach is that one can get diverse good tradeoff solutions, which are called the nondominated set, at a time. However, the users usually need one promising solution and not the entire set of nondominated solutions. Therefore, decision makers are needed to recommend the most promising solution among them. The decision criteria can vary according to the purpose of probe design, but we use a melting temperature variation as decision maker, as we mentioned in Section II-B. Although there are other features for controlling experimental conditions, melting temperature uniformity is one of the most important features. A melting temperature variation is measured by the standard deviation of the melting temperature of probes, which are estimated by the nearest neighbor model [14]. This criterion recommends the probe set, which has the minimum standard deviation of melting temperatures, and can be formulated as

$$\arg\min_{k} stdev\left(Tm\left(p_i^{(k)}, t_i\right)\right). \qquad (8)$$

Here, $stdev(\cdot)$ represents the standard deviation, $Tm(x,y)$ means the highest melting temperature for the most stable hybridization between sequences x and y, and $p_i^{(k)}$ denotes the ith probe in the kth set at the final archive produced by ϵ-MOEA.

IV. Experimental Result

To demonstrate the ability of EvoOligo, we design the probes for genomes of 19 different types of HPV and 52 genes in the Arabidopsis Calmodulin multigene family. The 19 HPV types, which have high potential to cause cervical cancer [23], are chosen by Biomedlab Co., Korea. The 52 Arabidopsis Calmodulin multigene family is composed of six AtCaMs (*Arabidopsis thaliana* Calmodulins) and 46 AtCMLs (*Arabidopsis thaliana* Calmodulinlike proteins) based on the recent study [24]. These genes function as sensors or regulators of the well-known key messenger "calcium (Ca^{2+})" in a network of signal transduction pathways to efficient adaptation [25]. Since the similarity among AtCaMs/AtCMLs sequences in each group is relatively high, care must be taken when designing probes for these groups of sequences.

The parameter settings for experiments are given as follows: The length of the probe is set to 30 nt for HPV and 24 nt for AtCaMs/AtCMLs. For the preprocessing step, we experimentally find the parameter setting that showed good performance: $(\alpha,\beta,\gamma)=(0.2,0.1,0.9)$ for HPV and $(\alpha,\beta,\gamma)=(0.1,0.0,0.9)$ for AtCaMs/AtCMLs. For the ϵ-multiobjective evolutionary algorithm, the crossover and mutation rates are set as conventional values of 0.9 and 0.01, respectively, and the ϵ is empirically set as 1. The population size is 50, and the maximum generation is 300. Except for the aforementioned parameters, the common default values are used for Primer3, ClustalW, Mfold, BLAT, BLAST, and NACST/Sim.

We performed various experiments to investigate the merits of EvoOligo. First, probe sets for HPV and AtCaMs/AtCMLs designed by EvoOligo are compared with the sets of the well-known probe design programs, such as OligoArray [2] and OligoWiz [7]. Then, the merits of evolutionary optimization of the probe are confirmed. Lastly, the effects of preprocessing are shown.

A. Criteria for Comparison of Probe Sets

A precise comparison of two sets of probes is a difficult problem. Although a real microarray experiment can be the most convincing approach, it heavily depends on the experimental condition. In literature, however, a few researchers empirically established the guideline for oligonucleotide probes of length 50 or 70 nt [8] and length 20 nt [9]. In these works, it has been shown that the free-energy values of target binding, nontarget binding, self-structure, and sequence similarity have some correlation with the probe's signal intensity on the microarray. Therefore, we choose seven criteria for probe reliability, as summarized in Tables I and II: # cross hybridization, BLAT search, BLAST match, ΔG_{25} (nontarget), ΔG_{25} (target), ΔG_{25} (self-structure), and Std_{Tm}.

The first four criteria are related to the specificity of probes. Both BLAT search and BLAST match calculate the specificity of probes by sequence similarity. While BLAT search calculates the similarity only within the given target sequences, BLAST match tests the similarity over the whole sequence database of the target organism. We used the BLAST program [26] with the database "nr" from NCBI (http://www.ncbi.nlm.nih.gov/) for HPVs and the TAIR database [27] for AtCaMs/AtCMLs. Based on the guideline suggested in [8], we regarded the probes whose sequence similarity with nontarget genes is larger than 85% as liable to cross hybridization. Both # cross hybridization and ΔG_{25} (nontarget) measure the specificity of the probes by calculating the free energy of nonspecific hybridization between the probes and the input sequences. ΔG_{25} (nontarget) measures the free energy for the most stable cross hybridization

TABLE I

CRITERIA FOR PROBE SET COMPARISON. MIN. AND MAX. MEAN THE MINIMIZATION AND MAXIMIZATION FUNCTIONS, RESPECTIVELY

Criteria	Formulation	Description	
BLAT search	$\sum_i \sum_{j \neq i} block\_count(p_i, t_j)$	Min.	Block count measure of sequence similarity among probes and non-target input sequences by BLAT.
BLAST match	$\sum_i I(identity(p_i, DB - \{t_i\}) > 0.85)$	Min.	Number of probes having sequence identity larger than 85% with non-target genes in the global database using BLAST.
# cross-hybridization	$\sum_i \sum_{j \neq i} I(Tm(p_i, t_j) > 37)$	Min.	Number of possible cross-hybridization by NACST/Sim.
ΔG_{25} (non-target)	$(\sum_i \min_{j \neq i} FEnergy(p_i, t_j))/n$	Max.	Average free energy for the most stable cross-hybridization by Mfold.
ΔG_{25} (target)	$(\sum_i FEnergy(p_i, t_i))/n$	Min.	Average free energy for binding to target by Mfold.
ΔG_{25} (self structure)	$(\sum_i FEnergy(p_i))/n$	Max.	Average free energy for the most stable self structure of probes by Mfold.
Std$_{Tm}$	$stdev(Tm(p_i, t_i))$	Min.	Standard deviation of melting temperatures of the probes by Mfold.

TABLE II

CLASSIFICATION OF CRITERIA FOR PROBE SET COMPARISON

	Specificity		Sensitivity	Uniformity
	Local	Global		
Sequence Similarity	BLAT search	BLAST match	ΔG_{25} (target) ΔG_{25} (self structure)	Std$_{Tm}$
Free Energy	# cross-hybridization ΔG_{25} (non-target)	Not Available		

TABLE III

HPV PROBE SET COMPARISON RESULTS BETWEEN EVOOLIGO, OLIGOARRAY, OLIGOWIZ, AND BIOMEDLAB. $_{\mathrm{Pre}}$ IS THE PROBE SET WITH PREPROCESSING, AND $_{\mathrm{WS}}$ IS THE PROBE SET WITH THE WHOLE SEQUENCES (WITHOUT PREPROCESSING). THE MEAN AND THE STANDARD DEVIATION (STD. DEV.) RESULTS OF ALL EVOOLIGO EXPERIMENTS ARE OBTAINED FROM 20 INDEPENDENT RUNS. THE RESULTS OF BIOMEDLAB ARE ADAPTED FROM [11]. THE MELTING TEMPERATURE IS CALCULATED BY THE NEAREST NEIGHBOR MODEL WITH 0.5-M Na$^+$ CONCENTRATION. THE CRITERIA ARE GROUPED USING THE CLASSIFICATION IN TABLE II. THE BEST RESULTS AMONG PROBE DESIGN TOOLS ARE MARKED IN BOLDFACE

	EvoOligo$_{Pre}$	EvoOligo$_{WS}$	OligoArray	OligoWiz	Biomedlab
	Mean ± Std. dev.	Mean ± Std. dev.			
BLAT search	$\mathbf{0.10 \pm 0.32}$	$\mathbf{0.10 \pm 0.32}$	2	1	0
BLAST match	8.90 ± 1.53	8.00 ± 2.11	**6**	9	5
# cross-hybridization	$\mathbf{141.20 \pm 26.69}$	160.90 ± 30.13	268	290	268
ΔG_{25} (non-target) (kcal/mol)	-23.22 ± 0.89	$\mathbf{-23.09 \pm 0.75}$	-25.93	-27.66	-21.86
ΔG_{25} (target) (kcal/mol)	-40.85 ± 0.69	-41.30 ± 0.80	**-44.79**	-43.76	-44.32
ΔG_{25} (self structure) (kcal/mol)	$\mathbf{-1.00 \pm 0.46}$	-1.12 ± 0.31	-2.47	-2.21	-2.02
Std$_{Tm}$	4.06 ± 0.63	4.15 ± 0.79	2.75	**1.28**	4.95

by Mfold [13]. On the other hand, # cross hybridization using NACST/Sim [28] checks all possible cross hybridizations between the probe and nontarget sequences. The possibility of cross hybridization is determined by its melting temperature using the nearest neighbor model [14]. If a cross hybridization has a higher melting temperature than a temperature specified by the experimental condition T, the cross hybridization could be possible. We set the room temperature to $T = 37$ °C, which is also the default condition when we calculate (4) and (5). The next two criteria are about the sensitivity of probes. ΔG_{25} (target) measures the free energies for the hybridization with target. In addition, ΔG_{25} (self-structure) measures the free energies for the secondary structure of the probes. In addition, the last criterion concerns the uniformity of the melting temperatures of probes.

B. Comparison With Other Tools

EvoOligo was compared with well-known programs, such as OligoArray [2] and OligoWiz [7]. OligoArray uses a simple filtering method based on sequence matching and free-energy calculation, whereas OligoWiz adopts dynamic programming with thermodynamic parameters. For fair comparison, we tried to set the parameters for these programs as close to EvoOligo as possible. Since OligoWiz and OligoArray use deterministic algorithms, both results show only one probe set each.

The comparison results of the final probe sets are shown in Tables III and IV. The results can be explained using the classification in Table II. The probe sets optimized by EvoOligo show better specificity for both gene groups: less number of cross hybridization, almost zero BLAT search, higher value for ΔG_{25} (nontarget), and similar BLAST match. From the aspect of specificity by free-energy calculation such as # cross hybridization and ΔG_{25} (nontarget), EvoOligo found much better probe sets for both HPV and AtCaMs/AtCMLs genes. In addition, EvoOligo found similar quality probe sets, compared with the previous tools, in the aspect of specificity by sequence similarity. Since EvoOligo tries to optimize probes using free-energy calculation, the results using the sequence similarity are not significantly improved. However, the results are still comparative. Generally, free-energy calculation is known to be more accurate than sequence similarity measure [14]; therefore, we can conclude that EvoOligo outperforms OligoArray and OligoWiz in probe specificity.

Regarding sensitivity, EvoOligo found poor probe sets in terms of ΔG_{25} (target). However, in terms of ΔG_{25}

TABLE IV
AtCaMs/AtCMLs PROBE SET COMPARISON RESULTS AMONG EVOOLIGO, OLIGOARRAY, AND OLIGOWIZ. THE BEST RESULTS AMONG PROBE DESIGN TOOLS ARE MARKED IN BOLDFACE

	EvoOligo$_{Pre}$	EvoOligo$_{WS}$	OligoArray	OligoWiz
	Mean ± Std. dev.	Mean ± Std. dev.		
BLAT search	0.10 ± 0.32	$\mathbf{0.00 \pm 0.00}$	1	1
BLAST match	6.00 ± 2.10	$\mathbf{4.40 \pm 1.35}$	5	6
# cross-hybridization	187.85 ± 24.21	$\mathbf{147.10 \pm 36.46}$	282	368
ΔG_{25} (non-target) (kcal/mol)	-19.79 ± 0.60	$\mathbf{-18.21 \pm 0.54}$	-20.51	-21.29
ΔG_{25} (target) (kcal/mol)	-34.66 ± 0.26	-33.65 ± 0.37	-35.66	**-35.99**
ΔG_{25} (self structure) (kcal/mol)	$\mathbf{-1.01 \pm 0.20}$	-1.03 ± 0.20	-1.39	-0.89
Std$_{Tm}$	4.29 ± 0.58	4.51 ± 0.64	2.97	**1.90**

(self-structure), EvoOligo found a better probe set than both tools for HPV and better probes than OligoArray and poorer ones than OligoWiz. Although EvoOligo showed slightly worse sensitivity, it still seems to be enough to form a stable hybridization with its target. Thus, EvoOligo found the reliable probe sets for both genes in probe sensitivity as well.

Lastly, EvoOligo produced poor probe sets for melting temperature uniformity. However, the bigger variance of melting temperatures could be explained by the difference of the probe search strategy. In OligoArray and OligoWiz, some candidate probes are removed to obtain a narrow range of melting temperature, and the surviving probes receive scores according to other criteria, which means that the criterion of uniform melting temperature was given priority over other criteria. On the other hand, EvoOligo does not consider the uniformity of melting temperatures in the evolutionary optimization step, yet the variance is still comparative.

In particular, EvoOligo found the comparative probe set, compared with the commercial Biomedlab probe set. Although the Biomedlab probe set was carefully designed by human experts, it does not outperform the EvoOligo probe set. However, EvoOligo found much better probes than Biomedlab for # cross hybridization, which may be the most important criterion among seven comparison criteria [10]. Additionally, since # cross hybridization checks the probability of cross hybridization between all possible probe–target pairs, it can show the advantage of optimizing combination of probes for targets over independently searching each probe. In terms of # cross hybridization, OligoArray and OligoWiz shows similar results as that of Biomedlab, and EvoOligo outperforms OligoArray, OligoWiz, and Biomedlab. This result is very promising and confirms the capability of EvoOligo with the evolutionary approach.

Tables III and IV also show the comparison results between EvoOligo with (EvoOligo$_{\mathrm{Pre}}$) and without preprocessing (EvoOligo$_{\mathrm{WS}}$). As shown in the tables, the selected probe candidate region affects the final probe reliability. For HPV, EvoOligo$_{\mathrm{Pre}}$ found a better probe set in terms of # cross hybridization, ΔG_{25} (self-structure), and Std$_{Tm}$. In other aspects, EvoOligo$_{\mathrm{Pre}}$ shows similar performance as that of EvoOligo$_{\mathrm{WS}}$. Therefore, the preprocessing step is helpful in finding better probe sets for HPV in terms of specificity, sensitivity, and melting temperature uniformity. When we focused on # cross hybridization, the necessity of the preprocessing is obvious. However, the preprocessing step does not show significant improvement for AtCaMs/AtCMLs. EvoOligo$_{\mathrm{WS}}$ outperformed EvoOligo$_{\mathrm{Pre}}$ in terms of specificity, whereas EvoOligo$_{\mathrm{Pre}}$ showed better results for sensitivity and uniformity. We suspect that the performance of EvoOligo$_{\mathrm{Pre}}$ for AtCaMs/AtCMLs is affected by the following reasons: First, the length of AtCaMs/AtCMLs genes is relatively short. The average length of AtCaMs/AtCMLs is 752.12 ± 250.71 nt, which is much shorter than that of HPV, which is 7873.83 ± 61.11 nt. Second, the similarity among AtCaMs/AtCMLs sequences is high. Therefore, it is hard to find the probe candidate area for AtCaMs/AtCMLs using the preprocessing step. Nonetheless, EvoOligo$_{\mathrm{Pre}}$ still found a better probe set than the previous tools. A more detailed explanation on preprocessing will be presented in Section IV-D.

C. *Effects of Evolutionary Probe Optimization*

We also investigated the merits of an evolutionary process for probe optimization. In our previous work [11], we noticed that an evolutionary process is an efficient way to find better probe sets. In addition, we also found that a small number of generations (1000 generations) showed remarkable improvement. To reconfirm these conclusions, we compared the results with 300 maximum generations, which is a much smaller number of generations than that in previous experiments.

Fig. 7 shows the change of the best # cross hybridization value in the archive over generations since # cross hybridization can be a representative criterion. The values are averaged over 20 independent runs using the same setting as in Tables III and IV. As evolution proceeds, probe sets are quickly improved. Within a short time of 150 generations, a twice better probe set can be found. Therefore, we can again conclude that the evolutionary process is effective for the probe selection procedure. Notice that we only used a relatively small size of population, i.e., 50. If we increase the population size, we can get more promising solutions. If users can bear the relatively long computation time, increasing the size of population can be an alternative to increasing the reliability of the final probe set.

D. *Effects of Preprocessing*

We looked into the detailed effects of preprocessing parameter setting to set the appropriate values for the parameters. Although parameters α, β, and γ in Fig. 5 can take any value between 0 and 1, some combinations of these values are invalid, depending on the target sequences. Under invalid settings, every primer pair generated by Primer3 is rejected,

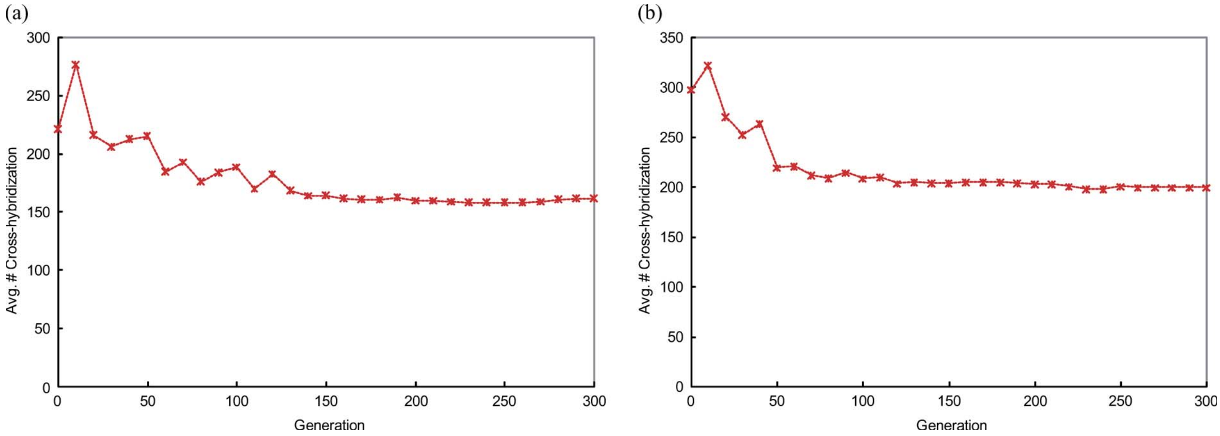


Fig. 7. Evolutionary optimization result. The average value is obtained by 20 independent experiments. (a) For HPV. (b) For AtCaMs/AtCMLs.

and the preprocessing step fails to select an appropriate probe region. We tested the following valid parameter values: for HPV, $\gamma = 0.9$, $\alpha \in \{0.05, 0.1, 0.15, 0.2\}$, and $\beta \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$, and for AtCaMs/AtCMLs, $\beta = 0$, $\alpha \in \{0.05, 0.1, 0.15, 0.2, 0.25\}$, and $\gamma \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$, except for the combination of $(\alpha, \gamma) = (0.05, 0.9)$.

The effect of choosing a candidate area is explained in Table V. We calculated # cross hybridization for the best probe set in each archive for the given parameters. As shown in the table, the parameter setting has a significant effect on the reliability of the final probes. The average # cross hybridizations of HPV vary from 126.60 to 175.60, and those of AtCaMs/AtCMLs vary from 192.00 to 297.00. Although we empirically selected the preprocessing parameters based on melting temperature uniformity, the selected parameter setting also showed the best performance in terms of # cross hybridization. Interestingly, the minimum of # cross hybridization in AtCaMs/AtCMLs genes is larger than the maximum of one in HPV genes, although HPV genes are almost ten times longer than AtCaMs/AtCMLs genes. This means that AtCaMs/AtCMLs gene sequences are very similar to each other. When we inspected the multiple alignments of genes, HPV showed more distinction between conserved and less-conserved regions than AtCaMs/AtCMLs. However, comparing Table V with Tables III and IV from the aspect of mean values of # cross hybridization, EvoOligo found better probes than the previous tools, except only one case in AtCaMs/AtCMLs. These results may imply that, if there are some conserved regions in the target sequences, the preprocessing step can successfully find those subsequence and improve the final probe quality.

Then, we investigated the computational overhead of the preprocessing step. The average computational time of the evolutionary optimization process is 10 024.2 and 39 560 s for AtCaMs/AtCMLs and HPV, respectively. The preprocessing time for AtCaMs/AtCMLs and HPV is 56.2 and 1845 s, respectively. The experiments were performed in a Linux machine with an Intel Pentium-4 3.00-GHz central processing unit and 2-GB memory. These results show that the preprocessing step requires 0.5% additional computational time for AtCaMs/AtCMLs and 4.66% for HPV. Therefore, the preprocessing step requires only allowable additional computational time, considering the improvement of probe quality.

TABLE V
# CROSS-HYBRIDIZATION RESULTS OF THE BEST PROBE SET IN EACH ARCHIVE FOR THE POSSIBLE PREPROCESSING STEP PARAMETER SETTINGS. WE RUN TEN TIMES FOR EACH EXPERIMENT. BOLD RESULTS REPRESENT THE PARAMETER SETTING USED IN TABLES III AND IV, AND CAN BE DISSIMILAR FROM THE RESULTS IN TABLES III AND IV DUE TO THE DIFFERENT NUMBER OF EXPERIMENTS

HPV			AtCaMs/AtCMLs		
α	β	# cross-hybridization (Mean ± Std. dev)	α	γ	# cross-hybridization (Mean ± Std. dev)
0.05	0.1	157.40 ± 27.12	0.05	0.1	251.00 ± 44.37
	0.2	166.60 ± 45.52		0.3	239.20 ± 41.94
	0.3	174.80 ± 31.44		0.5	225.20 ± 32.15
	0.4	161.80 ± 33.94		0.7	240.00 ± 18.64
	0.5	161.40 ± 36.48			
0.1	0.1	148.20 ± 31.71	0.1	0.1	216.60 ± 36.90
	0.2	164.00 ± 37.84		0.3	206.20 ± 16.54
	0.3	153.40 ± 22.31		0.5	226.00 ± 45.43
	0.4	175.60 ± 61.89		0.7	257.80 ± 56.51
	0.5	151.60 ± 17.10		0.9	$\mathbf{192.00 \pm 27.12}$
0.15	0.1	135.80 ± 32.65	0.15	0.1	225.20 ± 49.86
	0.2	162.00 ± 42.74		0.3	233.00 ± 39.21
	0.3	137.20 ± 14.65		0.5	221.40 ± 57.96
	0.4	135.60 ± 29.30		0.7	231.00 ± 39.39
	0.5	167.40 ± 9.13		0.9	199.00 ± 23.10
0.2	0.1	$\mathbf{125.60 \pm 8.65}$	0.2	0.1	297.00 ± 56.28
	0.2	157.00 ± 26.13		0.3	258.60 ± 26.86
	0.3	162.60 ± 42.90		0.5	255.20 ± 57.29
	0.4	143.60 ± 26.93		0.7	245.80 ± 15.74
	0.5	150.60 ± 15.99		0.9	200.20 ± 22.95
			0.25	0.1	252.60 ± 42.67
				0.3	232.80 ± 28.47
				0.5	228.00 ± 25.68
				0.7	199.40 ± 38.32
				0.9	222.60 ± 24.51

V. CONCLUSION

We have developed the EvoOligo platform to design an optimal probe set for the customized microarray experiments. To improve the reliability of the final probe set, EvoOligo incorporates ϵ-MOEA with constraint handling. The experimental results in designing probes for 19 HPV genomes and 52 Arabidopsis Calmodulin genes have proven the usefulness of the multiobjective evolutionary approach on the probe design. Compared with simple filtering methods and dynamic programming, the multiobjective evolutionary approach could improve the probe quality with small generations. The evolutionary search helped to find the optimized combination of probes for the given sequence sets, rather than to independently search for each target. In addition, we have tried to increase the reliability of the probe set by adopting the preprocessing step.

We have also focused on user accessibility of EvoOligo. Users can input the sequences into the EvoOligo system via a web browser, easily tune the necessary parameters, and get the optimal probe set by e-mail. Therefore, users do not need to install or manage the program.

The remaining work for this study will be automatic selection of parameters for preprocessing. Although preprocessing affects the final quality of probes, it requires many parameters that cannot easily be determined. Therefore, an automatic guide for these parameters will improve the usefulness of EvoOligo.

ACKNOWLEDGMENT

The authors would like to thank ICT, Seoul National University, which provided research facilities for this study.

REFERENCES

[1] P. M. K. Gordon and C. W. Sensen, "Osprey: A comprehensive tool employing novel methods for the design of oligonucleotides for DNA sequencing and microarrays," *Nucleic Acids Res.*, vol. 32, no. 17, p. e133, Sep. 2004.

[2] J.-M. Rouillard, M. Zuker, and E. Gulari, "OligoArray 2.0: Design of oligonucleotide probes for DNA microarrays using a thermodynamic approach," *Nucleic Acids Res.*, vol. 31, no. 12, pp. 3057–3062, Jun. 2003.

[3] X. Wang and B. Seed, "Selection of oligonucleotide probes for protein coding sequences," *Bioinformatics*, vol. 19, no. 7, pp. 796–802, May 2003.

[4] S. Drmanac, N. A. Stavropoulos, I. Labat, J. Vonau, B. Hauser, M. B. Soares, and R. Drmanac, "Gene-representing cDNA clusters defined by hybridization of 57 419 clones from infant brain libraries with short oligonucleotide probes," *Genomics*, vol. 37, no. 1, pp. 29–40, Oct. 1996.

[5] R. Herwig, A. O. Schmitt, M. Steinfath, J. O'Brien, H. Seidel, S. Meier-Ewert, H. Lehrach, and U. Radelof, "Information theoretical probe selection for hybridisation experiments," *Bioinformatics*, vol. 16, no. 10, pp. 890–898, Oct. 2000.

[6] F. Li and G. D. Stormo, "Selection of optimal DNA oligos for gene expression arrays," *Bioinformatics*, vol. 17, no. 11, pp. 1067–1076, Nov. 2001.

[7] R. Wernersson and H. B. Nielsen, "OligoWiz 2.0—Integrating sequence feature annotation into the design of microarray probes," *Nucleic Acids Res.*, vol. 33, pp. W611–W615, Jul. 2005. (Web Server issue).

[8] Z. He, L. Wu, X. Li, M. W. Fields, and J. Zhou, "Empirical establishment of oligonucleotide probe design criteria," *Appl. Environ. Microbiol.*, vol. 71, no. 7, pp. 3753–3760, Jul. 2005.

[9] O. V. Matveeva, S. A. Shabalina, V. A. Nemtsov, A. D. Tsodikov, R. F. Gesteland, and J. F. Atkins, "Thermodynamic calculations and statistical correlations for oligo-probes design," *Nucleic Acids Res.*, vol. 31, no. 14, pp. 4211–4217, Jul. 2003.

[10] C. Wu, R. Carta, and L. Zhang, "Sequence dependence of cross-hybridization on short oligo microarrays," *Nucleic Acids Res.*, vol. 33, no. 9, p. e84, May 2005.

[11] S.-Y. Shin, I.-H. Lee, and B.-T. Zhang, *Microarray Probe Design Using ϵ-Multi-Objective Evolutionary Algorithms With Thermodynamic Criteria*, vol. 3907. Berlin, Germany: Springer-Verlag, 2006, pp. 184–195.

[12] K. Deb, *Multi-Objective Optimization Using Evolutionary Algorithms*. New York: Wiley, 2001.

[13] M. Zuker, "Mfold web server for nucleic acid folding and hybridization prediction," *Nucleic Acids Res.*, vol. 31, no. 13, pp. 3406–3415, Jul. 2003.

[14] J. SantaLucia, Jr., "A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics," *Proc. Nat. Acad. Sci. U.S.A.*, vol. 95, no. 4, pp. 1460–1465, Feb. 1998.

[15] S.-Y. Shin, I.-H. Lee, D. Kim, and B.-T. Zhang, "Multiobjective evolutionary optimization of DNA sequences for reliable DNA computing," *IEEE Trans. Evol. Comput.*, vol. 9, no. 2, pp. 143–158, Apr. 2005.

[16] I.-H. Lee, S.-Y. Shin, and B.-T. Zhang, *Multiplex PCR Assay Design by Hybrid Multiobjective Evolutionary Algorithm*, vol. 4403. Berlin, Germany: Springer-Verlag, 2007, pp. 376–385.

[17] M. Laumanns, L. Thiele, K. Deb, and E. Zitzler, "Combining convergence and diversity in evolutionary multiobjective optimization," *Evol. Comput.*, vol. 10, no. 3, pp. 263–282, 2002.

[18] K. Deb, M. Mohan, and S. Mishra, "Towards a quick computation of well-spread Pareto-optimal solutions," in *Proc. 2nd Int. Conf. Evol. Multi-Criterion Optim.*, 2003, pp. 222–236.

[19] I.-H. Lee, S.-Y. Shin, and B.-T. Zhang, "Experimental analysis of ϵ- multiobjective evolutionary algorithm," in *Proc. 5th Int. Conf. Simul. Evol. Learn.*, 2004, p. SWP-1/127.

[20] R. Chenna, H. Sugawara, T. Koike, R. Lopez, T. J. Gibson, D. G. Higgins, and J. D. Thompson, "Multiple sequence alignment with the Clustal series of programs," *Nucleic Acids Res.*, vol. 31, no. 13, pp. 3497–3500, Jul. 2003.

[21] S. Rozen and H. Skaletsky, "Primer3 on the WWW for general users and for biologist programmers," *Methods Mol. Biol.*, vol. 132, pp. 365–386, 2000.

[22] W. J. Kent, "BLAT—The BLAST-like alignment tool," *Genome Res.*, vol. 12, no. 4, pp. 656–664, Apr. 2002.

[23] J. M. M. Walboomers, M. V. Jacobs, M. M. Manos, F. X. Bosch, J. A. Kummer, K. V. Shah, P. J. F. Snijders, J. Peto, C. J. L. M. Meijer, and N. Munoz, "Human Papillomavirus is a necessary cause of invasive cervical cancer worldwide," *J. Pathol.*, vol. 189, no. 1, pp. 12–19, Sep. 1999.

[24] E. McCormack, Y.-C. Tsai, and J. Braam, "Handling calcium signaling: Arabidopsis CaMs and CMLs," *Trends Plant Sci.*, vol. 10, no. 8, pp. 383–389, Aug. 2005.

[25] V. S. Reddy, G. S. Ali, and A. S. N. Reddy, "Genes encoding calmodulin-binding proteins in the Arabidopsis genome," *J. Biol. Chem.*, vol. 277, no. 12, pp. 9840–9852, 2002.

[26] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, "Gapped BLAST and PSI-BLAST: A new generation of protein database search programs," *Nucleic Acids Res.*, vol. 25, no. 17, pp. 3389–3402, Sep. 1997.

[27] S. Y. Rhee, W. Beavis, T. Z. Berardini, G. Chen, D. Dixon, A. Doyle, M. Garcia-Hernandez, E. Huala, G. Lander, M. Montoya, N. Miller, L. A. Mueller, S. Mundodi, L. Reiser, J. Tacklind, D. C. Weems, Y. Wu, I. Xu, D. Yoo, J. Yoon, and P. Zhang, "The Arabidopsis Information Resource (TAIR): A model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community," *Nucleic Acids Res.*, vol. 31, no. 1, pp. 224–228, Jan. 2003.

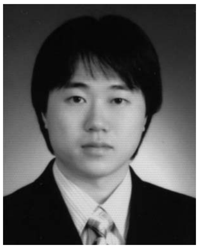
[28] S.-Y. Shin, H.-Y. Jang, M.-H. Tak, and B.-T. Zhang, "Simulation of DNA hybridization chain reaction based on thermodynamics and artificial chemistry," in *Prelim. Proc. 9th Int. Meeting DNA Based Comput.*, 2004, p. 451.

Soo-Yong Shin received the B.S., M.S., and Ph.D. degrees in computer science and engineering from Seoul National University, Seoul, Korea, in 1998, 2000, and 2005, respectively.

He is currently a Research Professor with the Medical Information Center, Seoul National University Hospital (SNUH), Seoul. Prior to joining SNUH, he was a Guest Researcher with the National Institute of Standards and Technology, Gaithersburg, MD, from 2006 to 2008. From March 2004 to August 2004, he was a Visiting Student with the Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge. His research interests include multiobjective evolutionary algorithms, biomolecular computing, medical informatics, and data mining.

In-Hee Lee received the B.S. degree in computer engineering from Seoul National University (SNU), Seoul, Korea, in 2001. She is currently working toward the Ph.D. degree with the Biointelligence Laboratory, School of Computer Science and Engineering, SNU.

Her research interests include multiobjective evolutionary algorithms, estimation of distribution algorithms, and DNA computing.



Young-Min Cho received the B.S. degree in computer engineering from Seoul National University (SNU), Seoul, Korea, in 2001. He is currently working toward the Ph.D. degree in computer science with the University of California San Diego, La Jolla.

His research interests include artificial intelligence, machine learning, and optimization.

Kyung-Ae Yang received the Ph.D. degree in molecular biology from Gyeongsang National University, Jinju, Korea, in 2005.

She is currently a Postdoctoral Researcher with the Center for Bioinformation Technology (CBIT), Seoul National University, Seoul, Korea. Her research interests include plant molecular biology, DNA computing, and bioinformatics.

Byoung-Tak Zhang received the B.S. and M.S. degrees in computer science and engineering from Seoul National University (SNU), Seoul, Korea, in 1986 and 1988, respectively, and the Ph.D. degree in computer science from the University of Bonn, Bonn, Germany, in 1992.

He is currently a Professor with the School of Computer Science and Engineering and the Graduate Programs in Bioinformatics, Brain Science, and Cognitive Science, SNU, and directs the Biointelligence Laboratory and the Center for Bioinformation Technology (CBIT). Prior to joining SNU, he was a Research Associate with the German National Research Center for Information Technology (GMD) from 1992 to 1995. From August 2003 to August 2004, he was a Visiting Professor with the Computer Science and Artificial Intelligence Laboratory (CSAIL), Massachusetts Institute of Technology (MIT), Cambridge. His research interests include probabilistic models of learning and evolution, biomolecular/DNA computing, and molecular learning/evolvable machines.

Prof. Zhang serves as an Associate Editor for the IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION, Advances in Natural Computation, and Genomics and Informatics. He serves on the Editorial Board of Genetic Programming and Evolvable Machines and Applied Soft Computing.