Neurocomputing 73 (2010) 2407–2418

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Optimal decoding and minimal length for the non-unique oligonucleotide probe selection problem

Laleh Soltan Ghoraie [a,*], Robin Gras [a], Lili Wang [b], Alioune Ngom [a]

[a] School of Computer Science, 5115 Lambton Tower, University of Windsor, 401 Sunset Avenue, Windsor, Ontario, Canada N9B 3P4
[b] School of Computing, 556 Goodwin Hall, Queen's University, Kingston, Ontario, Canada K7L 3N6

ARTICLE INFO

Available online 29 June 2010

Keywords:
Microarray
Probe selection
Target
Estimation of distribution algorithm
Bayesian optimization algorithm
Heuristic
Multiobjective optimization
Decoding
Markov chain Monte Carlo

ABSTRACT

One of the applications of DNA microarrays is recognizing the presence or absence of different biological components (*targets*) in a sample. Hence, the quality of the microarrays design which includes selecting short Oligonucleotide sequences (*probes*) to be affixed on the surface of the microarray becomes a major issue. A good design is the one that contains the minimum possible number of probes while having an acceptable ability in identifying the targets existing in the sample. This paper focuses on the problem of computing the minimal set of probes which is able to identify each target of a sample, referred to as *non-unique oligonucleotide probe selection*. We present the application of an *estimation of distribution algorithm* (*EDA*) named *Bayesian optimization algorithm* (*BOA*) to this problem, for the first time. The proposed approach considers integration of BOA and one simple heuristic introduced for the non-unique probe selection problem. The results provided by this approach compare favorably with the state-of-the-art methods in the single target case. While most of the recent research works on this problem has been focusing on the single target case only, we present the application of our method in integration with *decoding* approach in a multiobjective optimization framework for solving the problem in the case of multiple targets.

1. Introduction

Microarrays are tools used for performing many hybridization experiments in parallel. As noted by Schliep et al. [18], two main applications are considered for microarrays. A first application is measuring the expression levels of thousands of genes simultaneously. Gene expression level is measured based on the amount of mRNA sequences bound or hybridized to their complementary sequences affixed on the surface of the microarray. The complementary sequences are called *probes* which are typically short DNA strands about 8–30 bp [24]. The second important application of miocroarrays is the identification of unknown biological components in a sample [9]. Knowing the sequences affixed on the microarray and considering the hybridization pattern of a sample, one can infer which targets exist in the sample by observing appropriate hybridization reactions [18]. Finding an appropriate set of probes to be affixed on the surface of microarray, or in other words, finding a good *design* for microarray is a crucial task. The appropriate design should lead to cost-efficient experiments. Therefore, while the quality of the probe set is important, the objective of finding the minimal set of probes also should be considered. The quality of the probe set is discussed in terms of its ability to identify the unknown targets in the sample.

Two approaches are considered for the probe selection problem, namely, *unique* and *non-unique* probe selection. In the unique probe selection, for each single target there is one unique probe designed to hybridize only to the target. In this case, in specified experimental conditions, the probe should not hybridize to other targets except for its intended target. However, due to high levels of similarity in families of closely related gene sequences, finding unique probes for each target is almost impossible [9,10,13,16,18,22–24]. When many targets are similar, experimental errors increase. In these cases, an alternative approach is applying non-unique probes.

The non-unique probes are designed to hybridize to more than one target. The non-unique probe selection problem is to find the smallest probe set that is able to uniquely identify a set of targets in the sample [24]. Minimizing the probe set is a reasonable objective. Smaller microarray designs occupy less space on the surface of microarray. This leads to use smaller chips, and reduce the costs considerably [18].

Our focus in this paper is on the non-unique probe selection. We propose a method for solving the non-unique probe selection problem. Given a design containing candidate non-unique probes,

* Corresponding author.
E-mail addresses: soltanl@uwindsor.ca (L.S. Ghoraie), rgras@cs.uwindsor.ca (R. Gras), lili@cs.queensu.ca (L. Wang), angom@cs.uwindsor.ca (A. Ngom).

0925-2312/$ - see front matter Crown Copyright © 2010 Published by Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2010.02.026

our task is to analyze and minimize the design in order to select the best possible probe set. The initial design is presented as a target–probe incidence matrix. Target–probe incidence matrices contain the targets and probes and their hybridization patterns. The included probes are the high quality ones selected among all possible non-unique probes [9]. Computing the initial target–probe incidence matrix in not a trivial task [10].

Many parameters such as secondary structure, salt concentration, GC content, hybridization energy, etc., influence the quality of the probes hybridization [18], and should be considered carefully in selecting the candidate probes. For instance, at a given temperature and salt concentration, all probes should exhibit the same hybridization affinity [10]. Moreover, hybridization errors such as cross-hybridization, self-hybridization, and non-sensitive hybridization should also be taken into account in computing the set of candidate probes for the Oligonucleotide probe selection [23]. Candidate probes should neither be self-complementary nor should cross-hybridize [10].

It should be noted that we assume that the problem of computing the target–probe incidence matrix has been solved, and our focus is minimizing the design given by this matrix.

This paper is organized as follows. Section 2 provides a detailed description of the non-unique probe selection problem. Related work on non-unique probe selection is reviewed in Section 3. In Section 4, we present our approach for solving the non-unique probe selection problem. A review of the main concepts of Bayesian optimization algorithm (BOA) is presented in Section 5, and its advantages over the genetic algorithms (GA) are discussed. Also, the heuristic which we have integrated into the BOA is discussed in Section 6. At the end of this section, we explain how and why we integrate the heuristic into the BOA. The multiobjective optimization technique and decoding idea applied in this work are discussed in Sections 7 and 8, respectively. We discuss the results of our experiments in Section 9. In Section 10, complexity of the components of the approach is discussed. Finally, we conclude this research work with discussion of possible future research directions and open problems appears in Section 11.

2. Non-unique probe selection problem

We illustrate the probe selection problem with an example. Assume that we have a target–probe incidence matrix $H=(h_{ij})$ of a set of three targets $(t_1,\ldots,t_3)$ and five probes $(p_1,\ldots,p_5)$, where $h_{ij}=1$, if probe j hybridizes to target i, and 0 otherwise (see Table 1). The problem is to find the minimal set of probes which identifies all targets in the sample. First, we assume that the sample contains a single target. Using a probe set of $\{p_1, p_2\}$, we can recognize the four different situations of 'no target present in the sample', 't_1 is present', 't_2 is present', and 't_3 is present' in the sample. The minimal set of probes in this case is $\{p_1, p_2\}$ since $\{p_1\}$ or $\{p_2\}$ cannot detect these four situations.

Consider the case that multiple targets are present in the sample. In this case, the chosen probe set should be able to distinguish between the events in which all subsets (of all possible cardinalities) of the target set may occur. The probe set $\{p_1, p_2\}$ is not good enough for this purpose. With this probe set, we cannot recognize between the case of having subset $\{t_1, t_2\}$ and $\{t_2, t_3\}$ in the sample. However, the probe set $\{p_3, p_4, p_5\}$ can distinguish between all events in this case. It should be noted that the incidence matrix presented here is an unreal example, and its dimensions (number of probes and targets) are not representative of the real datasets of non-unique probe selection problem. For instance, the smallest incidence matrix in the literature contains about 256 targets and 2786 probes. For more information on the datasets properties, see Table 5.

Table 1
Sample target–probe incidence matrix.

	p_1	p_2	p_3	p_4	p_5
t_1	0	1	1	0	0
t_2	1	0	0	1	0
t_3	1	1	0	0	1

Now, a more formal definition of the probe selection problem is presented.

Given the target–probe incidence matrix H, and parameters $s_{min} \in \mathbb{N}$ and $c_{min} \in \mathbb{N}$, the goal is to select a minimal probe set such that each target is hybridized by at least c_{min} probes (*minimum coverage* constraint), and any two subsets of targets are separated by means of at least s_{min} probes (*minimum separation* constraint) [9,10].

A probe *separates* two subsets of targets if it hybridizes to exactly one of them. We say that a probe hybridizes to a set of targets when it hybridizes to at least one of the targets in the target set [18]. In other words, assume two target sets of S and T. If $P(S)$ and $P(T)$ are the set of probes hybridizing to S and T, respectively, a probe p separates these two sets of targets if $p \in P(S)\Delta P(T)$ [18]. Δ denotes symmetric set difference. Moreover, target sets S and T *are* s_{min}—separable if at least s_{min} probes separates them, that is $|P(S)\Delta P(T)| \geq s_{min}$.

The probe selection is proven to be a NP-hard problem [5], and is considered as a variation of the combinatorial optimization problem *minimal set covering problem*. We consider the problem in both cases of single target and multiple targets in the sample. The focus of the literature has mostly been on the problem in case of single target, although multiple targets in the sample are more realistic. In most of the real experiments of target–probe hybridization, several targets exist simultaneously in the sample, and in general, the identity of targets in the sample is unknown in advance. Therefore, it is crucial for the selected probe set of the final design to have the ability to identify several targets in the sample.

As mentioned, the problem can be approached as an optimization problem. The search space of the problem consists of 2^p (p=number of probes) possible solutions which makes this problem very difficult to solve exhaustively, even with powerful computers [16]. We propose a method based on an estimation distribution algorithms (EDA), named Bayesian optimization algorithm (BOA) integrated with a simple probe selection heuristic for both cases of single target and multiple targets in sample. This work is the first one which considers the ability of the probes to recognize multiple targets in the sample explicitly as an objective of the optimization algorithm.

3. Previous work

Most of the research works dedicated to the non-unique probe selection problem have focused on the case of single target in sample. As mentioned, this case is a simplified version of the non-unique probe selection problem. Rash and Gusfield [17] considered genes as strings and the probes as substrings of these original strings. They used suffix tree to identify the critical substrings and integer linear programming in order to solve the optimization problem. They applied CPLEX (an ILP solver) to solve the ILP problem. Schliep et al. [18] introduced a fast heuristic for minimizing the size of the probe set. Since guaranteeing the separation of all possible subsets of the original target set was computationally impossible by their heuristic, they could only

guarantee the separation of up to a randomly chosen number N (e.g. $N=500{,}000$) of pairs of target subsets. In this work, for the first time the idea of *decoding* was proposed. They presented a Bayesian method in order to evaluate the ability of the obtained probe set by their fast heuristic in identifying multiple targets. In this work, cross-hybridization and experimental errors were explicitly taken into account for the first time. Klau et al. [10] extended this work. CPLEX was applied to solve the ILP.

The ILP formulation extended to a more general version which also includes the group separation [9] in which groups correspond to multiple targets. They proposed a branch-and-cut algorithm to solve the ILP. By this extension, the assumption of the multiple targets was realized. Focusing on the single target case, Meneses et al. [13] used a two-phase heuristic including, first, construction of a feasible solution containing enough probes able to satisfy the constraints, and second, reducing the size of the probe set. Ragle et al. [16] applied a cutting-plane approach with reasonable computation time, and achieved the best results for some of the benchmark datasets in case of single target. Without using any *a priori* method to decrease the number of initial probes, the cutting-place algorithm relaxes a constraint set in order to find the lower bounds on the number of the required probes for an optimal solution. Then it improves the lower bound till an optimal solution is obtained. Wang and Ngom [23] presented deterministic heuristics in order to solve the ILP formulation, and reduce the size of final probe set. They applied their heuristic in order to introduce a population-based approach (without learning phase) for coverage and separation in order to guide the search for the appropriate probe set in case of single target in the sample. Recently, Wang et al. [22] presented a combination of the genetic algorithm and the selection functions used in [23], and obtained results which are in most cases better than results of Ragle et al. [16]. Finally, Soltan Ghoraie et al. [20], focusing on the single target case, proposed a compound method of Bayesian optimization algorithm (BOA) and a simple heuristic of Wang and Ngom [23] to solve the optimization problem of non-unique probe selection. This paper extends the mentioned approach to be able to solve the more realistic problem in case of presence of multiple targets in the sample.

4. Our approach

Our approach is based on integration of Bayesian optimization algorithm (BOA) (see Section 5) and a heuristic named dominated row covering (DRC) which was proposed in [23] for solving the problem of non-unique probe selection (see Section 6). The non-unique probe selection problem has been considered as an optimization problem for the cases of single target and multiple targets in the sample. We approach the problem in case of single target in the sample as a one-objective optimization problem. The objective is minimizing the number of selected probes.

The case of multiple targets in the sample is considered as a two-objective optimization problem. While first objective is minimizing the probe set, the other one is the ability of the selected set in identifying a predetermined number of targets in the sample. Several methods have been proposed to solve multiobjective optimization problems efficiently by means of evolutionary-based algorithms such as BOA (see Section 7). We have applied one of the most efficient methods proposed in the literature.

The definition of the non-unique probe selection problem is realistic when the possibility of presence of a set of targets in the sample is considered. Only in this case, the obtained solutions are practical solutions. Therefore, evaluating the ability of the selected (by means of any method) probe sets in identifying targets of the sample is a critical task. Our work is the first one that explicitly seeks to maximize the ability of a probe set in identifying multiple targets in the sample, along with the goal of minimizing the probe set. In order to measure the ability of selected probe set in identifying multiple targets, we have applied *decoding* idea proposed by Schliep et al. [18] (see Section 8).

5. Bayesian optimization algorithm

Estimation of distribution algorithm (EDA) method was introduced by Larranaga and Lozano [12] and Mühlenbein and Paaß [14]. EDAs are also called probabilistic model-building genetic algorithms (PM-BGA) which extend the concept of classical GAs. Targeting more efficient exploration of the search space, EDA approach has been proposed. In EDA optimization methods, a sample of the search space is generated and the information extracted from that sample is used in order to explore the search space more efficiently.

The EDA (Algorithm 1) is an iterative approach. In initialization (1), a set of random solutions is generated which is the first sample of search space (2); the quality of solutions is evaluated (3); a subset of high quality solutions that have more probability to be chosen is selected (4); a probabilistic model of the sample is constructed, and the model is used to generate a new set of solutions (5). The algorithm is repeated from evaluation step.

Algorithm 1. EDA

1. (Random) initialization of set of solutions S_0
2. $S=S_0$
3. Evaluation of S
4. Select set of promising solutions S_l
5. Build probabilistic model M of S_l
6. Sample from the Model M and generate new set of solutions S
7. Repeat from 3.

In BOA, which was first proposed by Pelikan [15], the constructed probabilistic model is a Bayesian network. A Bayesian network can be considered as a directed acyclic graph in which the nodes represent the variables of the problem, and the directed edges introduced between some nodes represent the dependencies among the variables. The important advantage of constructing a Bayesian network is discovering and representing the possible dependencies between the variables of the problem. The discovered dependencies which are extracted from the sample of search space, are used to accomplish the target of BOA to explore the search space more efficiently.

Based on the generic algorithm of EDA, in BOA, a probabilistic model which is a Bayesian network is constructed in step (5). Learning a Bayesian network is basically a two-step process. First the dependencies should be discovered which means an appropriate network structure should be found, and second, the conditional probabilities between the variables should be estimated. A local search algorithm is used for the problem of building the best network from the sample in each iteration of BOA. A metric to measure the quality of the built network directs the local search. For further information on building Bayesian networks, see [8]. After constructing the network, the joint probabilities should be estimated. These probabilities can be estimated based on the frequency of occurrences of the variables in the sample.

In optimization problems, there is a difficult class of problems which contain dependencies among variables, and classical GAs has been shown not to be able to solve these problems properly [6]. On the other hand, BOA approach has been more successful in solving such problems. We are interested in applying BOA

approach for the complex problem of nonunique probe selection optimization problem. In this problem, we considered that each binary variable represents the presence or absence of a particular probe in the final design matrix. The dependencies among variables represent the fact that choosing a particular probe have a consequence on the choice of other probes in an optimal solution. Pelikan and Goldberg [4,15] have proven that when the number of variables and the number of dependencies are n and k, respectively, the size of the sample should be about of $O(2^k \cdot n^{1.05})$ to guarantee convergence.

There are several advantages in applying this new approach. First, BOA is known as an efficient way to solve complex optimization problems. Therefore, it is interesting to compare it with other methods applied to the non-unique probe selection problem. Second, EDA methods, by working on the samples of the search space and deducing the properties of dependencies among the variables of the problem, are able to reveal new knowledge about the biological mechanisms involved. Finally, with the study of the results obtained from experimenting different values of the parameter k, BOA provides the ability to evaluate the level of complexity of the non-unique probe selection in general, and the specific complexity of the classical set of problems applied to evaluate the algorithms used for solving this problem in particular.

6. Dominated row covering heuristic

As mentioned above, our algorithm integrates a simple heuristic to the BOA. The heuristic dominated row covering (DRC) was proposed by Wang and Ngom [23]. Given the target–probe incidence matrix H, probe set $P=\{p_1,\ldots,p_n\}$, and the target set $T=\{t_1,\ldots,t_m\}$, two main functions $C(p_j)$ (*coverage function*) and $S(p_j)$ (*separation function*) have been defined for this heuristic as follows:

$$C(p_j) = \max_{t_i \in T_{p_j}} \{cov(p_j, t_i) \mid 1 \le j \le n\} \quad (1)$$

where T_{p_j} is the set of targets covered by p_j

$$S(p_j) = \max_{t_{ik} \in T^2_{p_j}} \{sep(p_j, t_{ik}) \mid 1 \le j \le n\} \quad (2)$$

where $T^2_{p_j}$ is the set of target pairs separated by probe p_j.

Function C favors the selection of probes that c_{min}—cover *dominated targets*. Target t_i dominates target t_j, if $P_{t_j} \subset P_{t_i}$. Function S favors the selection of the probes that s_{min}—separate *dominated target pairs*. Target pair t_{ij} dominates target pair t_{kl}, if $P_{t_{ij}} \subset P_{t_{kl}}$. The functions $C(p_j)$ *and* $S(p_j)$ have been defined as the maximum between the values of the function *cov* and *sep*, respectively.

The functions *cov* and *sep* have been defined over $P \times T$ and $P \times T^2$, respectively, as follows:

$$sep(p_j, t_{ik}) = |h_{ij} - h_{kj}| \frac{s_{min}}{|P_{t_{ik}}|}, \quad p_j \in P_{t_{ik}}, \quad t_{ik} \in T^2 \quad (3)$$

$$cov(p_j, t_i) = h_{ij} \frac{c_{min}}{|P_{t_i}|}, \quad p_j \in P_{t_i}, \quad t_i \in T \quad (4)$$

where P_{ti} is the set of probes hybridizing to target t_i, and P_{tik} is the set of probes separating target-pair t_{ik}.

Value of $sep(p_j, t_{ik})$ is what p_j can contribute to satisfy the separation constraint for target-pair t_{ik}. Value of $cov(p_j, t_i)$ is the amount that p_j contributes to satisfy the coverage constraint for target t_i. Hence, S and C are the maximum values that p_j can contribute to satisfy the minimum separation and coverage constraints, respectively.

The selection function $D(p_j)$ which has been defined as follows will indicate the degree of contribution of p_j to the minimal solutions

$$D(p_j) = \max\{C(p_j), S(p_j) \mid 1 \le j \le n\} \quad (5)$$

The probes with high value of $D(p_j)$ are *good* probes that will be selected for the solution probe set. The coverage and separation functions of DRC have been calculated for the target–probe incidence matrix of Table 2, in Tables 3 and 4, respectively [23].

The DRC algorithm consists of three phases of: *Initialization, Construction*, and *Reduction* (*see* Algorithm 2).

In the initialization phase, the $D(p)$ value is computed for each probe of the original probe set. The probes for which $D(p)=1$ are added to an initial probe set (P_{ini}). This probe set is most probably a non-feasible solution. Therefore, in the construction phase (Algorithm 2), high-degree (high-value in D) probes are added to the initial probe set repeatedly until we obtain a feasible solution (P_{con}). In the Reduction phase (Algorithm 2), the low-degree (low-value in D) probes are removed repeatedly, as long as, the feasibility of the solution is not disturbed. At the end of this phase, we may obtain a minimal feasible solution (P_{red}).

Algorithm 2. Dominated row covering heuristic

Input: $T=\{t_1,\ldots,t_m\}$, $P=\{p_1,\ldots,p_n\}$, and $H=[h_{ij}]$
Output: Near-minimal solution P_{min}
Begin
/*__Initialization Phase__*/
Compute $D(p)$ for all $p \in P$ using Eqs. (8)–(10)
$P_{ini} \leftarrow \{p \in P | D(p)=1\}$/* *essential probes only**/
/*__Construction Phase__*/
$P_{sol} \rightarrow P_{ini}$
Sort $P \setminus P_{sol}$ in decreasing order of $D(p)$
for *each target* t_i *not* c_{min}*—covered by* P_{sol} **do**
$n_i \leftarrow$ #probes needed to complete c_{min}-coverage of t_i
$P_{sol} \leftarrow P_{sol} \cup U_{l=1}^{l=n_i}$ {next highest-degree probe $p_l \in P \setminus P_{sol}$ that covers t_i}

Table 2
Target–probe incidence matrix.

	p_1	p_2	p_3	p_4	p_5	p_6
t_1	1	1	0	1	0	1
t_2	1	0	1	0	0	1
t_3	0	1	1	1	1	1
t_4	0	0	1	1	1	0

Table 3
Coverage function table: C has been calculated based on the DRC definition.

	p_1	p_2	p_3	p_4	p_5	p_6
t_1	$c_{min}/4$	$c_{min}/4$	0	$c_{min}/4$	0	$c_{min}/4$
t_2	$c_{min}/3$	0	$c_{min}/3$	0	0	$c_{min}/3$
t_3	0	$c_{min}/5$	$c_{min}/5$	$c_{min}/5$	$c_{min}/5$	$c_{min}/5$
t_4	0	0	$c_{min}/3$	$c_{min}/3$	$c_{min}/3$	0
C	$c_{min}/3$	$c_{min}/4$	$c_{min}/3$	$c_{min}/3$	$c_{min}/3$	$c_{min}/3$

Table 4
Separation function table: S has been calculated based on the DRC definition.

	p_1	p_2	p_3	p_4	p_5	p_6
t_{12}	0	$s_{min}/3$	$s_{min}/3$	$s_{min}/3$	0	0
t_{13}	$s_{min}/3$	0	$s_{min}/3$	0	$s_{min}/3$	0
t_{14}	$s_{min}/5$	$s_{min}/5$	$s_{min}/5$	0	$s_{min}/5$	$s_{min}/5$
t_{23}	$s_{min}/4$	$s_{min}/4$	0	$s_{min}/4$	$s_{min}/4$	0
t_{24}	$s_{min}/4$	0	0	$s_{min}/4$	$s_{min}/4$	$s_{min}/4$
t_{34}	0	$s_{min}/2$	0	0	0	$s_{min}/2$
S	$s_{min}/3$	$s_{min}/2$	$s_{min}/3$	$s_{min}/3$	$s_{min}/3$	$s_{min}/2$

end
for *each target-pair* t_{ik} *not* s_{min}*-separated by* P_{sol} **do**
$n_{ik} \leftarrow$ #probes needed to complete s_{min}-separation of t_{ik}
$P_{sol} \leftarrow P_{sol} \cup U_{l=1}^{l=n_{ik}}$ {next highest-degree probe $p_l \in P \setminus P_{sol}$ that separates t_{ik}}
end
/*__Reduction Phase__*/
$P_{min} \leftarrow P_{sol}$
$H \leftarrow H|_{P_{min}}$, /*restriction of H to probes in P_{min}*/
Compute $D(p)$ for all $p \in P_{min}$
Sort $P_{del} \leftarrow \{p \in P_{min} | D(p) < 1\}$ in increasing $D(p)$
if $P_{min} \setminus \{p\}$ *is feasible for each* $p \in P_{del}$ **then**
$P_{min} \leftarrow P_{min}\{p\}$
end
Return final P_{min}
end

6.1. The combination of BOA and DRC

As mentioned, we have applied the modified version of BOA to the non-unique probe selection problem. In this version, we have integrated BOA with the DRC heuristic described above. The minimum set of probe should satisfy the coverage and separation constraints. Since the probe set found by BOA does not guarantee the constraints satisfaction, we have applied DRC heuristic in order to guarantee this issue. As mentioned, in each iterative step of BOA, a sample of solutions is generated. Each solution is a string of 0 and 1 which represents a set of probes. Each position in the string represents the presence or absence of a probe in the solution which is noted by 1 or 0, respectively.

After generating the sample of solutions, the feasibility of each solution should be guaranteed by computing the DRC heuristic. Hence, every solution generated by BOA in the current sample, is transformed by applying the heuristic, in order to respect the coverage and separation constraints.

In order to apply the Bayesian optimization algorithm, the objective(s) to be optimized should be determined. An objective is a function that measures the quality of the solutions for the given problem, and this measure will help explore the search space efficiently in order to find good solutions that optimize the objective. In single target case, the goal is minimization of the probe set. In multiple targets case, in addition to this goal, we want to maximize the ability of the found probe set in identifying several targets in the sample. These can be defined as the objective(s) for the BOA. Therefore, for the first goal, we use inverse of the length of a solution as our objective function. The length of a solution corresponds to the cardinality of probe set, and it is given by the number of ones in the solution. For the second goal, in the multiple targets case, we use a modified version of the *decoding* idea (see Section 8).

7. Multiobjective optimization

Multiobjective optimization refers to optimization problems with several separate objectives [1]. In these problems, each solution has a value for each objective. In other words, each solution has several objective values. The immediate problem caused by this property is how to judge about the overall fitness of solutions. For instance, a solution may have good values for some of the objectives, and have weak values for other objectives. Another solution may have average values for all the objectives. Which of these solutions is better? This major problem, especially cause the evolutionary-based optimization algorithms to be confused in convergence to the optimal solution [1]. There is no clear way to compare the quality of the solutions in this case. In such cases, a classical approach is to make a weighted sum over all the objectives and try to make a single compound objective to be able to judge about the overall fitness of the solutions. There are two major problems for this approach. First, finding the appropriate weights for each objective is not a trivial problem itself. Assigning wrong weights may cause the evolutionary-based algorithm to converge to an unacceptable solution. Second, sometimes assigning weights to separate objectives and combining them is as meaningless as comparing very different criteria and trying to judge which is better than the other. The literature approach this problem as a ranking problem, and different methods are proposed and examined in order to solve this problem.

In solving the non-unique probe selection problem in multiple targets case, we consider two major objectives. First objective is minimizing the cardinality of the probe set. Second one is maximizing the ability of recognizing multiple targets existing in the sample by selecting the most appropriate probes. These two objectives are somewhat contradictory. We know that in case of selecting more probes, the ability of probe set in recognizing the targets in the sample increases. Therefore, we decided to use one of the multiobjective optimization approaches for solving this problem, instead of combining these two objectives and making one single objective.

Bentley and Wakefield [1] have mentioned an important property for an appropriate ranking method for evaluating the solutions in multiobjective optimization problems. This property is range-independent. In most of the complex multiobjective problems, each objective has an effective range, and the function ranges is non-commensurable [19]. As a result, in case of combining different objectives and making one single objective from them, it is possible that the compound fitness is influenced by the values of the objectives of a larger range more than the objectives of smaller ranges. Hence, in order to ensure that all the objectives are treated equally, either all the objective ranges should be the same in order to make them commensurable, or the method should ensure that objectives are not directly compared with each other.

Bentley and Wakefield [1] have proposed six ranking methods for multiobjective optimization problems: three range-dependent and three range-independent. The most important one is weighted average ranking (WAR). In this method, the fitness values of the solutions for each objective are extracted and listed separately. The lists are sorted, and based on the position of each fitness value, a rank is assigned to the fitness value of the solution. For each solution, different ranks obtained by sorting each list of objectives is averaged. Since each objective has been treated separately, this method is range-independent.

Corne and Knowles [2] have evaluated seven ranking methods using a multiobjective evolutionary algorithm in cases of having 5, 10, 15, and 20 objectives. They have shown that the method of average ranking AR (modified version of the WAR of Bentley and Wakefield) outperforms the other algorithms in most cases. Based on their results, they recommended using this method for the 2–5 objectives problem.

We have applied this method in our experiments of two-objective problem for solving the non-unique probe selection problem in the multiple targets case. By applying multiobjective optimization technique with BOA, we have provided a framework for the problem of non-unique probe selection. New objectives for the problem which result from further studies based on the nature of the problem can be added to the framework easily.

8. Decoding

The *decoding* method proposed by Schliep et al. [18], uses a Bayesian framework to infer the presence of the targets in the sample. The method is based on Monte Carlo Markov Chain sampling and it explicitly allows for experimental errors. Assume a probe set of $\{p_1,\ldots,p_n\}$ as the solution of non-unique probe selection, and a result vector $r=(r_1,\ldots,r_n)$ in which each r_i corresponds to the result of hybridization (0 or 1) of the current sample of targets to the probe p_i. Given the mentioned result vector, the posterior probability that a set of targets S includes all the targets present in the sample is calculated by

Bayes formula as follows:

$$P[S|r]=\frac{P[r|S]P[S]}{P[r]} \tag{6}$$

$P[r|S]$ is the probability of observing the result vector r, while all and only targets of set S are present in the sample. In order to formulate the $P[r|S]$, two assumptions were made. First, the probability of observing a specific result is only related to the number of targets from the set S that a probe binds to. Second, the observed binding results of probes are independent from each other. Based on these assumptions, Schliep et al. [18] have defined the $P[r|S]$ as

$$P[r|S]=\prod_{p_j} f(r_j,|S(j)\cap S|) \tag{7}$$

where $S(j)$ is the set of targets probes p_j hybridizes to and $|S(j)\cap S|$ is the number of targets probe p_j hybridizes to and also are in the target set S. Note that r_j is either 0 or 1. $f(0,0)$, $f(0,\geq 1)$, $f(1,0)$, and $f(1,\geq 1)$ are different cases that this function will have. Considering f_p and f_n as false positive and false negative rates of the target–probe hybridization experiments, four cases of f, mentioned above, were set to $1-f_p$, f_n, f_p, and $1-f_n$, respectively.

A prior probability ($P[S]$) is assigned to every set S from the set of all subsets of the original targets set. This is the probability of finding only the targets of set S in the sample. The prior probability of observing k different targets in a sample is denoted by c_k, and the abundance of each target t_i in samples including more than one target is denoted by f_i. Hence, the prior probability has been defined as

$$P[S]\propto c_{|S|}\prod_{t_i\in S} f_i\prod_{t_i\notin S}(1-f_i) \tag{8}$$

In the non-unique probe selection, we are interested in calculating the probability of presence of target t in the sample, given the result vector r. This can be shown by the marginal $p[t_i|r]$ which can be calculated by the posterior of set S over all sets T that include targets t

$$P[t_i|r]\propto \sum_{S:t\in S} P[S|r] \tag{9}$$

Since $P[r]$ is not available, the posterior cannot be computed directly. On the other hand, computing the above equation requires an exponential time in terms of the number of targets. Therefore, the proposed method for this problem by Schliep et al. [18] is Markov Chain Monte Carlo. By sampling a sufficient number of sets S_k, the marginal $P[t_i|r]$ can be estimated as the frequency of observing t in the sets S_k. A Markov chain is constructed over all possible sets S, which is the space of all subsets of the original target set. By choosing $P[S|r]$ as the stationary distribution, Gibbs sampling is applied in this approach. The Markov chain is guaranteed to converge to a stationary distribution. After convergence, the relative frequency of the targets t_i in the states S_k that chain visited is used in estimation of the marginals $P[t_i|r]$.

The decoding software was provided to us by Dr. Schliep. We changed the software in order to use the decoding as one of our objectives in the optimization problem. In order to measure the ability of each probe set, obtained by BOA, in identifying a set of targets in the sample, we select a set of true targets. We introduce the experimental errors to the model. This also helps in solving the non-unique probe selection problem more realistically. The probes that hybridize to the true targets are assumed to be true positives. In experiments, we considered $f_n=0.05$ and $f_p=0.05$. We removed probes from the positive true probes according to the false positive rate, and also add probes to the positive probes set according to the false negative rate.

The obtained design (probe set) is the input for the decoding software, and the output is a ranked list of targets based on the probability of their presence in the sample. We examine the ranked list in order to find the true targets among them. We assume that a given set of targets are carefully identified if in the ranked list of all targets predicted by the decoding algorithm, the true targets existing in the sample are the only ones ranked in the first top positions. Based on this, we defined the decoding related objective for BOA.

In our experiments, we randomly select a subset of the original target set as the true targets set. For l randomly selected targets, there are l possible top positions of 0,1,2,…, $l-1$. We search the sorted list of targets produced by the decoding algorithm for the l true targets, and their positions. Hence, we will obtain a list of positions: $pos_1,pos_2,\ldots,pos_l$. The objective Obj_{dec} is defined as following:

$$Obj_{dec}=\frac{1}{\sum_{i=1}^{l} pos_i} \tag{10}$$

Hence, the maximum value for this objective happens when all the true targets are ranked in the top l position of the list. In this case, the summation is calculated as: $((l-1)\times l)/2$. We examine at most 100 targets of the sorted list. In case of not finding the true targets in the sorted list, their position value is set to 100. Therefore, the maximum value for the positions summation, which corresponds to the minimum value for the objective, is equal to: $l\times 100$. In this case, none of the initial true targets are found in the ranked list of the targets.

9. Results of computational experiments

We combined BOA with DRC heuristic for solving the non-unique probe selection problem for both cases of single target and multiple targets in the sample. In the single target case, as presented in [20], the results of applying our method indicated that we are able to improve the results obtained by the best methods in literature.

We have extended our method, using a multiobjective optimization technique, in order to cover the multiple targets case which is a more realistic problem. Since our method is basically a time-consuming one, we have applied message passing interface (MPI) technique [7] in order to decrease the execution time of the program. The MPI is a library of methods for distributed computing. It should be noted that since microarray design is not a repetitive task, the execution time of the method used for obtaining a good design is not important. Hence, different methods applied for the problem have been compared based on the cardinality of the final obtained probe set, and not the computational time. The experiments were written in c++ and conducted on Sharcnet systems [26].

The parameters of coverage and separation constraints (c_{min} and s_{min}) were set to ten and five, respectively. We calculated the appropriate sample size by applying the condition of convergence

for the BOA which was mentioned in Section 5. While n is the number of variables, the sample size should be of $O(2^k n^{1.05})$. The number of variables is equal to the number of real and virtual probes for each dataset in this problem. The virtual probes are added to the datasets to guarantee the feasibility of the original probe set. The feasibility is defined in terms of satisfying the coverage and separation constraints.

In all the experiments, we set the variable k to two. This parameter is equal to the maximum number of incoming edges to each node of the Bayesian Network used in the BOA software [27] to model every sample of the search space. Other parameters of BOA software have been set to their default values. For instance, the percentage of the offspring and parents in the sample was set to 50.

9.1. Data sets

We have performed the experiments on ten artificial datasets called $a_1,\ldots, a_5, b_1,\ldots,b_5$, and two real datasets *HIV*1 and *HIV*2. All previous studies mentioned in Section 3 have been conducted on the same datasets, except for the *HIV*1, and *HIV*2 that have not been used in [9,10]. As mentioned, the datasets are the target–probe incidence matrices. Properties of the datasets are presented in Table 5. Along with this information, the number of virtual probes required for each dataset has been noted.

Single target in sample the results of applying this approach for the case of single target are presented in Table 6. For this case, as mentioned in the [20], BOA was executed for 100 iterations. We have noted our approach by BOA+DRC, and compared it to the other methods in literature. Comparison is based on the number of probes in the final solution obtained by each method. Three columns have been included related to experiments performed by state-of-the-art approaches integer linear programming (ILP) [9,10], optimal cutting plane algorithm (OCP) [16], and genetic algorithm (DRC-GA) [22].

The illustrated results in Table 6 for algorithms DRC-GA and BOA+DRC are the best obtained results among several runs of the program. Since these optimization methods are evolutionary-based ones, and randomization is used in construction of the first sample of solutions, the programs should be run several times.

Table 7 contains information regarding five runs of BOA+DRC on the non-unique probe selection problem in case of single targets in the sample, and the minimum, average, and the maximum of the obtained results are noted by *min*, *ave*, and *max* in the table, respectively.

As shown in Table 6, the results obtained by the Ragle et al. [16] (noted as OCP) are considered as the best ones in the literature for the non-unique probe selection problem. On the other hand, the results of Wang et al. [22] (noted as DRC-GA) are comparable to (and in most cases better than) [16].

We have been able to improve the result of non-unique probe selection for dataset *HIV*2, and obtain the shortest solution length of 474. The results we obtained for datasets a_1, a_2, a_4, and *HIV*1 are also equal to the best results calculated for these datasets in the literature. Table 8 summarizes and presents comparison between our method and each of the three mentioned.

Table 5
Properties of the datasets used for experiments.

Set	$\|T\|$	$\|P\|$	$\|V\|$
a_1	256	2786	6
a_2	256	2821	2
a_3	256	2871	16
a_4	256	2954	2
a_5	256	2968	4
b_1	400	6292	0
b_2	400	6283	1
b_3	400	6311	5
b_4	400	6223	0
b_5	400	6285	3
*HIV*1	200	4806	20
*HIV*2	200	4686	35

The first ten are artificial, and the last two ones are real. Number of targets, probes, and virtual probes are noted by ($|T|$), ($|P|$), and ($|V|$), respectively.

Table 6
Comparison of the cardinality of the minimal probe set for different approaches: Performance of various algorithms evaluated using ten datasets with different number of targets ($|T|$), probes ($|P|$), and virtual probes ($|V|$).

Set	ILP [9,10]	OCP [16]	DRC-GA [22]	BOA+DRC [20]
a_1	503	509	502	502
a_2	519	494	490	490
a_3	516	543	534	533
a_4	540	539	537	537
a_5	504	529	528	528
b_1	879	830	839	834
b_2	938	842	852	846
b_3	891	827	835	829
b_4	915	873	879	875
b_5	946	874	890	879
*HIV*1	–	451	450	450
*HIV*2	–	479	476	474

Column DRC-GA and BOA+DRC contain the least obtained cardinalities in several runs.

9.2. Multiple targets in sample

As mentioned, we have extended our method to cover the case of multiple targets for the non-unique probe selection problem. We applied the multiobjective optimization technique presented in Section 7, and measured the ability of the probe set in

Table 7
Minimum, average and maximum cardinality of obtained probe sets by optimization algorithm over five runs of program.

Set	BOA+DRC		
	Min.	Ave.	Max.
a_1	502	502 ± 0	502
a_2	490	502 ± 0	490
a_3	533	533.4 ± 0.49	534
a_4	537	537 ± 0	537
a_5	528	528 ± 0	528
b_1	834	835.8 ± 0.98	837
b_2	846	848.6 ± 1.61	851
b_3	829	831.6 ± 1.38	834
b_4	875	876.4 ± 1.36	879
b_5	879	881.2 ± 1.6	883
*HIV*1	450	450.2 ± 0.4	451
*HIV*2	474	474.4 ± 0.41	475

Table 8
Comparison between BOA+DRC and ILP, OCP, and DRC-GA: Number of datasets for which our approach has obtained results better or worse than or equal to methods ILP, OCP, and DRC-GA.

	Worse	Equal	Better
ILP	2	0	8
OCP	5	0	7
GA-DRC	0	5	7

identifying a predetermined number of random targets in the sample as the second objective for our optimization problem. This ability was measured by applying the decoding idea described in Section 8.

The experiments were conducted in two main series of identification of five and ten targets, and identification of 15 and 20 targets in the sample. All experiments were performed while the number of generations for BOA was set to 40, and the BOA was combined with only the DRC heuristic in these experiments. Regarding the decoding program, we have set the number of warm-up steps and total steps 5000 and 50,000, respectively [18].

9.2.1. Identification of five and ten targets

In the first series of experiments, the goal was set to measure the ability of the solutions in identifying five and ten targets in the sample. The results are presented in Table 9. It should be mentioned that the cardinality of probe sets presented in this table are the best ones in three runs of the program for each dataset.

In the first step of experiments, we chose to measure the ability of the solutions in identifying five random targets in the sample. Investigating the obtained results, we realized that the identification ability of the solutions are higher than the expectation, and a randomly selected probe set (in first iteration of BOA) is able to identify five targets in the sample for all the datasets.

As presented in Table 9, the length of the minimal solutions (or number of probes in final probe sets) for all datasets are greater than what we achieved in one-objective optimization problem (Table 6). This is expected in multiobjective optimization. The optimization algorithm should compromise between optimizing each of the two objectives. Therefore, this is natural that objective length has not been minimized as before, especially while the two objectives are in contradiction with each other. As mentioned, a larger set of probes results in better decoding ability.

In next step, we decided to increase the number of the targets to ten in order to make a more difficult optimization problem. Even in this case, our observation was similar to the previous step regarding objectives length (cardinality of the probe sets) and decoding ability.

As mentioned before, we have set the separation constraint (s_{min}) to five. By applying the DRC heuristic (6) in our method, we guarantee the separation of all pairs of targets by at least five probes. Enforcing this constraint improves the decoding ability of the obtained probe sets by our method; But the number of targets that can be identified by the probe sets is not known and should be investigated. Therefore, by performing the mentioned experiments in case of five and ten targets in the sample, in fact, we determined the number of targets that can be identified by the probe sets obtained by our method.

We assumed that the problem of decoding could be modified to discovering a threshold for the difficulty of decoding for each dataset. That is, we can examine further in order to find the maximum number of multiple targets that can exist in the sample, and the solutions generated by our method can identify them properly. Finding this threshold and increasing it will make difficult enough optimization problems. We expect to obtain larger sets of probes by solving these optimization problems, as the reason was explained; But the obtained probe sets will have the ability of identifying larger numbers of targets in the sample which will be more realistic. We conducted another series of experiments to investigate our assumption more carefully (see Section 9.2.2).

9.2.2. Identification of 15 and 20 targets

Since the obtained probe sets by our method had a high ability to identify multiple targets in the sample, we tried to increase the number of targets in the sample, make a more difficult optimization problem and find the difficulty threshold of decoding problem for each dataset. Therefore, we examined the problem in case of 15 and 20 targets in the sample.

We conducted new experiments for all the datasets. Our observation about the cardinality of the obtained probe set in these two cases (of 15 and 20 targets) was the same as the cases of five and ten targets, that is, here, the obtained probe sets by multiobjective optimization are larger than the obtained probe sets by one-objective optimization problem, too.

Tables 10 and 11 contain the maximum, minimum, and average decoding scores obtained by three runs of our optimization program for all the datasets in cases of 15 and 20 targets in the sample, respectively.

Our observations of decoding ability of the probe sets were interesting. We realized that our attempt to find a difficulty threshold for the decoding problem was right. Not only we could find this threshold for some datasets, but also, by applying our proposed multiobjective framework, we could improve the decoding ability of the probe sets significantly. For instance, the improvements of the decoding score (in case of 15 targets) in 40 iterations of BOA for dataset a_3 is shown in Fig. 1.

In this figure, the maximum decoding score obtained in each iteration of BOA is presented. The maximum possible decoding score for the case of 15 targets is obtained when all the targets are identified by the probe set as the top 15 positions. Therefore, the value of the maximum score is $(1/105) \approx 0.009524$. As shown in

Table 9
Cardinality of minimum probe set obtained by applying the BOA+DRC in case of multiple targets in the sample-two cases of five and ten targets in the sample were considered.

Set	BOA+DRC (5 targets)	BOA+DRC (10 targets)
a_1	508	515
a_2	494	502
a_3	537	545
a_4	540	546
a_5	533	539
b_1	867	879
b_2	883	897
b_3	864	872
b_4	891	912
b_5	920	938
*HIV*1	456	458
*HIV*2	483	487

Table 10
Minimum, average, and maximum of decoding score obtained by three runs of the optimization program for the case of 15 targets in the sample.

Set	Dec score(15 targets)		
	Min.	Ave.	Max.
a_1	0.005235	$0.005235 \pm 0.5e-6$	0.005236
a_2	0.005235	$0.006578 \pm 20.22e-4$	0.009259
a_3	0.009434	$0.009470 \pm 3.8e-5$	0.009523
a_4	0.005236	$0.008062 \pm 199.9e-5$	0.009523
a_5	0.005235	$0.0066\,1 \pm 236.9e-5$	0.009346
b_1	0.009449	$0.009498 \pm 4.3e-5$	0.009523
b_2	0.009327	$0.009458 \pm 11.3e-5$	0.009523
b_3	0.009422	$0.009489 \pm 5.8e-5$	0.009523
b_4	0.009327	$0.009458 \pm 11.3e-5$	0.009523
b_5	0.009274	$0.009440 \pm 14.4e-5$	0.009523
*HIV*1	0.003597	$0.004095 \pm 74.4e-5$	0.004950
*HIV*2	0.005236	$0.007959 \pm 235.8e-5$	0.009346

Table 11
Minimum, average and maximum cardinality of obtained probe sets by optimization algorithm over three runs of program for the case of 20 targets in the sample.

Set	Dec score(20 targets)		
	Min.	Ave.	Max.
a_1	0.002488	$0.002579 \pm 1.19e-4$	0.002747
a_2	0.002695	$0.002743 \pm 2.8e-7$	0.002793
a_3	0.002824	$0.002829 \pm 1.4e-5$	0.002833
a_4	0.002808	$0.003148 \pm 3.70e-4$	0.003663
a_5	0.002293	$0.002518 \pm 25.4e-5$	0.002793
b_1	0.002391	$0.003343 \pm 161.5e-5$	0.005208
b_2	0.002789	$0.003252 \pm 45.1e-5$	0.003690
b_3	0.004356	$0.000508 \pm 494.3e-5$	0.005236
b_4	0.003798	$0.004700 \pm 83.8e-5$	0.005263
b_5	0.002785	$0.003210 \pm 45.5e-5$	0.003690
*HIV*1	0.001963	$0.002023 \pm 5.3e-5$	0.002062
*HIV*2	0.002564	$0.002683 \pm 10.4e-5$	0.002732

Table 12
Comparing the average decoding score (ave. decoding score) of the optimal probe set obtained by one-objective optimization with the maximum decoding score (max. decoding score) obtained by the multiobjective optimization in case of 15 targets in the sample.

Set	Ave dec. score	Ave. target position	Max. dec. score	Ave. target position
a_1	0.001300	51.28	0.005236	12.73
a_2	0.001304	51.12	**0.009259**	**7.2**
a_3	0.001335	49.93	**0.009523**	**7**
a_4	0.001338	49.82	**0.009523**	**7**
a_5	0.001218	54.73	**0.009346**	**7.13**
b_1	0.001499	44.47	**0.009523**	**7**
b_2	0.001486	44.86	**0.009523**	**7**
b_3	0.001477	45.14	**0.009523**	**7**
b_4	0.001627	40.97	**0.009523**	**7**
b_5	0.001476	45.17	**0.009523**	**7**
*HIV*1	0.000956	69.73	0.004950	13.46
*HIV*2	0.001196	55.74	**0.009346**	**7.13**

The average target position (ave. target position) corresponding to each score is also presented. Maximum possible decoding score (0.009523) has been obtained for almost all the datasets except for a1 and HIV1.

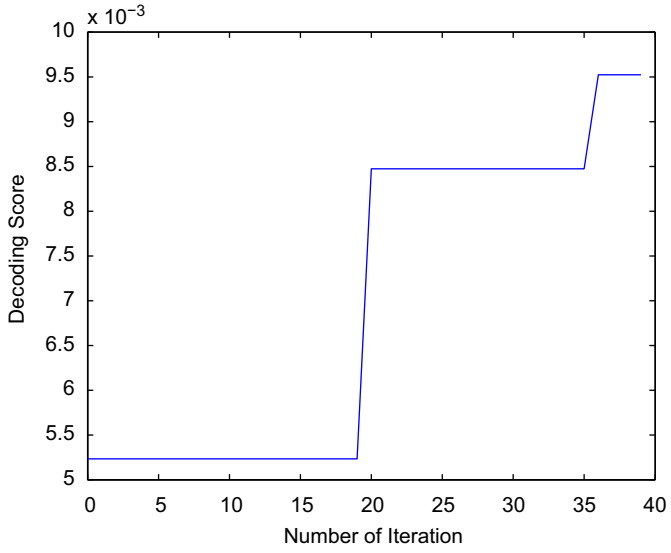


Fig. 1. Maximum decoding score for dataset a3 in 40 iterations of multiobjective optimization in case of 15 targets in the sample.

the figure, the maximum decoding score in iterations has been improved from 0.005235 to the maximum possible decoding score 0.009523. This indicates that our method has been able to solve this optimization problem quiet efficiently.

As described in Section 8, the inverse of the maximum decoding score in case of 15 targets $(\frac{1}{0.009524} \approx 105)$ is the summation of the targets positions. Therefore, $(\frac{105}{15} \approx 7)$ indicates the average targets positions in the optimal case. By inversing the decoding score, and dividing it by the number of targets in the sample, we calculate the average targets position corresponding to the decoding score

$$\text{Average targets position} = \frac{\sum_{i=1}^{l} pos_{t_i}}{l}, \quad 1 \le i \le l \tag{11}$$

where t_i is the target existing in the sample, and l is the number of targets in the sample.

The average targets position can be used for comparing the obtained results by different experiments. In order to show the targets identification improvements obtained by the multiobjective method, we calculated the decoding score for the optimal probe sets obtained by one-objective optimization problem (see Section 9.1), and averaged the score over 50 runs for each of the five datasets (a_1,…,a_5). We compared the calculated score with the maximum score obtained by multiobjective optimization. In all cases, considerable improvements were noticed. The scores and their associated average target position are demonstrated in Table 12. For instance, the average target position identified by the optimal probe set obtained in case of single target in sample, for dataset a_3, is 49.93. By applying multiobjective optimization method, we have improved this value to its best possible value (7) in case of 15 targets in the sample.

It should be noted that although the decoding ability of the probe sets has been significantly improved comparing with the probe sets obtained in single target case, the decoding score has not been improved considerably, during 40 iterations, for the dataset a1. The problem of identifying 15 targets in the sample can be considered a difficult problem for this dataset, and further attempts are required in order to solve this problem more efficiently.

The same calculations can be conducted for the case of 20 targets in the sample (see Table 13). The maximum decoding score in this case is $\frac{1}{190} \approx 0.005263$. 190 which is the summation of 20 targets positions results in $\frac{190}{15} \approx 12.67$ average target position for this case.

As presented in Table 13, comparing with the optimal probe set obtained by the one-objective optimization, probe set obtained by two-objective framework has higher ability in identification of targets. The maximum decoding score after 40 iterations of two-objective method is always greater than the average score calculated for the optimal solution obtained by one-objective optimization.

Since the optimization problem in case of 20 targets is a difficult problem, we did not notice a significant improvement in the value of decoding objective during the 40 iterations of our method for any of the datasets. It means that the current configuration of BOA is not able to solve this problem efficiently. Therefore, we should try to find a better BOA configuration to solve this case more efficiently. The possible modifications can be performed on the number of iterations of BOA.

On the other hand, we think that we should investigate the impact of the parameter 'maximum incoming edges' on the decoding objective. The maximum incoming edges (see Section 5), determines the level of dependency among variables in BOA.

Table 13
Comparing the average decoding score (ave. decoding score) of the optimal probe set obtained by one-objective optimization with the maximum decoding score (max. decoding score) obtained by the multiobjective optimization in case of 20 targets in the sample.

Set	Ave. dec. score	Ave. target position	Max. dec. score	Ave. target position
a_1	0.000920	54.35	0.002747	18.20
a_2	0.000898	55.68	0.002793	17.90
a_3	0.000885	56.50	0.002833	17.65
a_4	0.000988	50.61	0.003663	13.65
a_5	0.000828	60.39	0.002793	17.90
b_1	0.000989	50.56	**0.005208**	**9.6**
b_2	0.001067	46.86	0.003690	13.55
b_3	0.001177	42.48	**0.005236**	**9.54**
b_4	0.001152	43.40	**0.005263**	**9.5**
b_5	0.001037	48.22	0.003690	13.55
*HIV*1	0.000677	73.85	0.002062	24.24
*HIV*2	0.001134	44.09	0.002732	18

The average target position (ave. target position) corresponding to each score is also presented. The maximum possible decoding score (0.005263) has been obtained for dataset $b4$ and almost $b1$ and $b3$.

Table 14
Comparing cardinality of the minimum probe set obtained by one-objective optimization problem and the cardinality of the solution with the maximum decoding score in case of 20 targets in the sample.

Set	Minimum length (single target in sample)	Length (of the solution with maximum decoding score)
a_3	533	618
b_1	834	968
b_2	846	989
b_3	829	932
b_4	875	1159
b_5	879	1010
*HIV*1	450	525
*HIV*2	474	584

9.2.3. Comparison between optimized and random solutions of same length

Following the experiments illustrated in Section 9.2.2, we performed another series of interesting experiments on the dataset a_3, all the datasets of *b*-series, and *HIV*-datasets.

As mentioned before, the minimal length of solutions or the cardinality of the minimal probe set obtained by our multi-objective optimization framework is more than the minimal length obtained by the one-objective optimization approach. Furthermore, the solution with the minimal number of probes is not necessarily the one with the best decoding score.

In Table 14, the minimum length obtained in case of single target in the sample (experiments of Section 9.1 and Table 6) for some datasets are demonstrated. Along with these, the length of the solution with the maximum decoding value in case of 20 targets in the sample is indicated for mentioned datasets in Table 14. It should be mentioned that for both cases of single and multiple targets in the sample, the best (minimum) obtained probe set cardinality among several experiments has been provided.

We conducted a new comparison to illustrate the efficiency of our approach, as follows. We chose the minimum set of probes obtained by the one-objective optimization approach for each dataset, and added random probes to this set as far as constructing a set of the same cardinality mentioned in the third column of Table 14. Then, the decoding score of the resulted probe set, for each dataset, was compared with the obtained maximum decoding score in the case of 20 targets. The result is illustrated in Table 15.

Table 15
Comparing the decoding ability of the optimized solution in case of 20 targets in the sample to the decoding ability of a random solution of the same length.

Set	Random solution	Optimized solution
a_3	0.000869	0.002833
b_1	0.000893	0.005208
b_2	0.000909	0.003690
b_3	0.001047	0.005236
b_4	0.001094	0.005263
b_5	0.001010	0.003690
*HIV*1	0.000674	0.002062
*HIV*2	0.000778	0.002732

As noted in Table 15, in the second column, decoding score of a random solution of the same length of the optimal solution obtained by our two-objective framework is illustrated. In the third column, the maximum decoding value obtained for the case of 20 targets in the sample is shown. Comparing these two values for each dataset, it can be noticed that a considerable increase has been obtained by applying the optimization algorithm.

As mentioned before, by increasing the number of the probes, the decoding ability of the probe set also increases; We noticed that by increasing the cardinality of the probe set, the decoding ability did not increase as much as when we apply our optimization algorithm. This proved the efficiency of our algorithm from another aspect.

10. Complexity

Among the works on the non-unique probe selection problem in the literature, we notice [3,21] which have discussed the problem complexity theoretically, and do not contain and illustrate any results obtained by an implementation of a specific method.

The three mentioned works have been focusing on the minimization problem resulted from investigating the non-unique probe selection problem, and applying group testing approach in order to solve it. In this case, the minimization problem is considered as finding a d-disjunct submatrix of a given binary matrix. The submatrix should contain the same number of the columns, but the minimum number of rows.

According to the group testing approach, the columns of the matrix are same as the targets in non-unique probe selection problem, and the rows are the *pools* which are same as the probes

which hybridize to the targets. The definition of this binary matrix corresponds to the definition of the target–probe incidence matrix.

In a binary d-disjunct matrix, any union of the d columns cannot contain the $(d+1)$th column. Wang et al. in [21] have proved and concluded that, first, minimum d-disjunct submatrix problem (MIN-d-DS) is polynomial-time solvable in the case that all pools have size two, and second, this problem is MAX SNP-complete in the case that all pools have size at most two, and there is a polynomial-time approximation with performance ratio $1+2/(d+1)$ for $d \geq 1$.

Wang et al. also prove that if the all pools are of size of more than two, MIN-d-DS is still NP-hard, approximations with better performance may exist.

Cheng et al. also in [3] discusses the complexity of the non-unique probe selection based on the minimal d-separable matrix problem, which is proven to be *DP*-complete. A generalized decision version of the minimum d-separable submatrix problem is the d-separable submatrix with reserved rows, and can be solved by a nondeterministic machine. This problem is $\sum_2^p$-complete.

The base of our optimization framework is the EDA of BOA. The other major components of the presented framework in this paper are: the DRC heuristic and the decoding program. These include the time-consuming parts of the framework. Regarding determining the complexity of the whole framework, complexity of all the components are discussed.

Pelikan [15] discusses the complexity of BOA. The most dominant and time-consuming part of the algorithm is its network construction. As mentioned before, we have always performed the experiments with the parameter maximum incoming edges $k \geq 1$. In this case, the overall complexity of constructing Bayesian network is $O(k2^k n^2 N + kn^3)$ in which n is number of variables of the problem or the number of probes, and N is the number of the instances of solutions (sample) generated in each iteration of the algorithm. The value of this parameter was discussed in Section 5, based on the convergence condition of BOA. Assuming that k is a constant, the complexity can be rewritten as: $O(n^2 N + n^3)$.

The complexity of the DRC heuristic has been discussed by Wang et al. in [25]. The dominant part of this algorithms is the Construction phase, and results in complexity of $O(n^2 m^2)$ in which m is the number of targets and $n \geq m$.

Knill et al. have analyzed the decoding idea based on the *Markov Chain Monte Carlo* method in [11]. The described software in this work is the base of the decoding program we have applied in our framework. The complexity of decoding a result is proportional to number of steps $\times\, m \times d$. d is the average number of pools or probes each target hybridize to. As a routine in the MCMC methods, user can choose the number of warm-up steps and also the number of steps between states that are used for marginals calculations [18].

This analysis indicates that the most dominant component of our framework from running-time aspect is related to the BOA algorithm.

11. Conclusions (and future research)

In this paper, we extended the new approach proposed in [20] in order to solve the non-unique probe selection in multiple targets in the sample. In case of single target, our method which is a combination of an EDA named BOA and the heuristic DRC obtains results which compare favorably with the state-of-the-art. This method has proved its efficiency compared with other methods.

The case of multiple targets in the sample was specifically the subject of discussion in this paper. Most recent works have been focusing on the single target case in the sample, and ignoring the assumption of multiple targets in the sample. For this case, we applied an extended version of the combination of BOA and DRC [20], and considered a second objective for the problem which was ability of identification of multiple targets in the sample. By applying a modified version of the decoding, we tried to measure the second objective for solutions. We approached this problem as a two-objective optimization problem. Our method is the first one which proposes a multi-objective optimization framework which considers the decoding ability of the obtained probe sets along with the objective of cardinality of the probe sets.

Experiments in case of five and ten targets in the sample were conducted. Examining the results, we realized that identification of five or ten targets is not a difficult problem for the obtained probe sets. The separation constraint (s_{min}) in the non-unique probe selection problem improves the decoding ability of the obtained solutions (probe sets) by our method. Even in first iteration of the algorithm, we can find probe sets that are able to identify five or ten targets in the sample properly.

Since the ability of the solutions obtained by BOA+DRC in identifying the five and ten targets in the sample was already high, we investigated this problem for finding the maximum number of targets that can be identified by the solutions obtained by our method, and improving the ability of decoding. Assumption of 15 and 20 targets in the sample constructed difficult optimization problems. Our method was successful in solving the optimization problem for the case of 15 targets for almost all the datasets except for a_1 and *HIV*1, and for the case of 20 targets for dataset M and almost b_3 and b_1. Optimization led to obtaining maximum possible decoding ability for the probe sets after 40 iterations.

On the other hand, comparing the decoding ability of the probe sets obtained by one-objective and two-objective optimization, we noticed a significant improvement by applying two-objective framework for both cases of 15 and 20 targets in the sample. Also comparison between the decoding ability of the optimized probe sets and random probe sets of the same length also illustrated the efficiency of this method.

Our experiments were conducted while the parameter k of BOA software was set to two. In future, the experiments can be performed to investigate the impact of increase in parameters such as number of iterations of BOA and level of dependency among variables (k).

Moreover, we believe that our multiobjective-based method makes a flexible framework for the problem of non-unique probe selection. In further studies, it will be interesting to consider new objectives for this problem. For instance, the cost associated to adding a probe to a microarray chip may differ for several probes. Therefore, a third objective of obtaining the least expensive design can be considered for the problem. By applying our proposed approach, it will be possible to embed the new objectives to the problem by using current flexible structure proposed in this paper.

Acknowledgments

This work is supported by the NSERC grant ORGPIN 341854, the CRC grant 950-2-3617 and the CFI grant 203617. The authors gratefully acknowledge the comments and software TCPD (a modified version of MCPD [28]) provided by Dr. Alexander Schliep and Mr. Ole Schulz-Trieglaff that resulted in improvement of this investigation. We are also thankful of Dr. Pardalos and Dr.

Ragle who kindly provided us the data required for the experiments of the paper.

References

[1] P.J. Bentley, J.P. Wakefield, Finding acceptable solutions in the Pareto-optimal range using multiobjective genetic algorithms, in: Second Online World Conference on Soft Computing in Engineering Design and Manufacturing (WSC2) 5, 1998, pp. 242–249.

[2] D.W. Corne, J.D. Knowles, Techniques for highly multiobjective optimization: some non-dominated points are better than others, GECCO (2007) 773–780.

[3] Y. Cheng, K. Ko, W. Wu, On the complexity of non-unique probe selection, Theoretical Computer Science (2008) 120–125.

[4] D.E. Goldberg, The Design of Innovation: Lessons from and for Competent Genetic Algorithms, Kluwer Academic Publishers, 2002.

[5] M. Garey, D. Johnson, Computers and Intractability: A Guide to the Theory of NP-completeness, W. Freeman, San Francisco, 1979.

[6] R. Gras, How efficient are genetic algorithms to solve high epistasis deceptive problems?, in: IEEE Congress on Evolutionary Computation, Hong Kong, China, June 1–6, 2008, pp. 242–249.

[7] W. Gropp, E. Lusk, A. Skjellum, Using MPI: Portable parallel programming with the message-passing interface, MIT Press in Scientific and Engineering Computation Series, MA, USA, 1994.

[8] D. Heckerman, D. Geiger, D. Chickering, Learning Bayesian networks: the combination of knowledge and statistical data, in: Tenth Conference on Uncertainty in Artificial Intelligence, Seattle, WA, 1994, pp. 293–301.

[9] G.W. Klau, S. Rahmann, A. Schliep, M. Vingron, K. Reinert, Integer linear programming approaches for non-unique probe selection, Discrete Applied Mathematics 155 (2007) 840–856.

[10] G.W. Klau, S. Rahmann, A. Schliep, M. Vingron, K. Reinert, Optimal robust non-unique probe selection using integer linear programming, Bioinformatics 20 (2004) i186–i193.

[11] E. Knill, A. Schliep, D. Torney, Interpretation of pooling experiments using the Markov Chain Monte Carlo Method, Journal of Computational Biology 3 (1996) 395–406.

[12] P. Larranaga, J.A. Lozano, Estimation of Distribution Algorithms: A New tool for Evolutionary Computation, Kluwer Academic Publishers, 2001.

[13] C.N. Meneses, P.M. Pardalos, M.A. Ragle, A new approach to the non-unique probe selection problem, Annals of Biomedical Engineering 35 (4) (2007) 651–658.

[14] H. Muhlenbein, G. Paaß, From recombination of genes to the estimation of distributions I. Binary parameters, in: 4th International Conference on Parallel Problem Solving from Nature, September 22–26, 1996, pp. 178–187.

[15] M. Pelikan, Bayesian optimization algorithm: from single level to hierarchy, Ph.D. Thesis, University of Illinois, 2002.

[16] M.A. Ragle, J.C. Smith, P.M. Pardalos, An optimal cutting-plane algorithm for solving the non-unique probe selection problem, Annals of Biomedical Engineering 35 (11) (2007) 2023–2030.

[17] S. Rash, D. Gusfield, String barcoding: uncovering optimal virus signatures, in: Annual Conference on Research in Computational Molecular Biology, 2002, pp. 254–261.

[18] A. Schliep, D.C. Torney, S. Rahmann, Group testing with DNA chips: generating designs and decoding experiments, in: IEEE Computer Society Bioinformatics Conference (CSB'03), 2003, pp. 84–91.

[19] J.D. Schaffer, Multiple objective optimization with vector evaluated genetic algorithms, in: Genetic Algorithms and their Applications: 1st International Conference on Genetic Algorithms, 1985, pp. 93–100.

[20] L. Soltan Ghoraie, R. Gras, L. Wang, A. Ngom, Bayesain optimization algorithm for the non-unique oligonucleotide probe selection problem, in: 4th IAPR International Conference, Pattern Recognition in Bioinformatics (PRIB 2009), September 7–9, Sheffield, UK, pp. 365–376.

[21] F. Wang, H. Due, X. Jia, P. Deng, Non-unique probe selection and group testing, Theoretical Computer Science (2007) 29–32.

[22] L. Wang, A. Ngom, R. Gras, Non-unique oligonucleotide microarray probe selection method based on genetic algorithms, in: 2008 IEEE Congress on Evolutionary Computation, Hong Kong, China, June 1–6, 2008, pp. 1004–1010.

[23] L. Wang, A. Ngom, A model-based approach to the non-unique oligonucleotide probe selection problem, in: Second International Conference on Bio-Inspired Models of Network, Information, and Computing Systems (Bionetics 2007), Budapest, Hungary, December 10–13, 2007, ISBN: 978-963-9799-05-9.

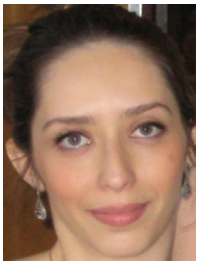
[24] L. Wang, A. Ngom, R. Gras, L. Rueda, Evolution strategy with greedy probe selection heuristics for the non-unique oligonucleotide probe selection problem, in: 2008 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB 2008), 2008, pp. 54–61.

[25] L. Wang, A. Ngom, R. Gras, L. Rueda, An evolutionary approach to the non-unique oligonucleotide probe selection problem, in: Springer Transactions on Computational Systems Biology, Lecture Notes in Bioinformatics, vol. 5410(X), December 2008, pp. 143–162.

[26] http://www.sharcnet.ca/.

[27] http://www.cs.umsl.edu/pelikan/software.html.

[28] http://algorithmics.molgen.mpg.de/Software/MCPD/.



Laleh Soltan Ghoraie is a M.Sc. candidate at University of Windsor. She has started her graduate studies under supervision of Dr. Robin Gras, in Fall 2007. Her research interest is Bioinformatics and Machine Learning. She received her B.Eng. from Shahid Beheshti University, Iran in 2005. She received her Masters degree in Computer Science at University of Windsor, Canada, in 2009, and started PhD program at University of Waterloo, Canada, in 2010.

Dr. Robin Gras is Associate Professor and Canadian Research Chair in Probabilistic Heuristics and Bioinformatics at the School of Computer Science of the University of Windsor. He is also cross-appointed by the Biological Department at the University of Windsor. He was senior scientist, from 2000 to 2004, in the Swiss Institute of Bioinformatics, Geneva, Switzerland after being post-doctorant from 1998 to 2000 in the same institute and lecturer, in 1998, at the University of Rennes, France. He received his B.Sc. and his M.Sc. in computer science at the University of Rennes. He completed his Ph.D. in computer science applied to bioinformatics at INRIA of Rennes in 1997, and obtained his Habilitation a Diriger la Recherche in 2004 in the University of Rennes. From 2000 to 2002 he was also consultant for GeneProt Inc. concerning the automation of protein identification and characterization process.

Lili Wang is a Ph.D. candidate at the School of Computing, Queen's University, Kingston, Canada. Her research interest is Bioinformatics. She received her M.Sc. from University of Windsor, Canada in 2008, and B.Sc. from Academy of Armoured Forces Engineering, Beijing, China in 2002.

Alioune Ngom has a B.Sc. in Computer Science from the University du Quebec a Trois-Rivieres (Trois-Rivieres, Quebec, Canada, in 1992) under the supervision of Dr. Corina Reischer, and a M.Sc. and Ph.D. in Computer Science from the University of Ottawa (Ottawa, Ontario, Canada, in 1995 and 1998), under the supervision of Dr. Ivan Stojmenovic. He joined the School of Computer Science at the University of Windsor in 2000 as an Assistant Professor, and have been appointed since 2004 as an Associate Professor. He was appointed as an Assistant Professor at the Department of Mathematics and Computer Science at Lakehead University (Thunder Bay, Ontario, Canada), 1998–2000, prior to joining the University of Windsor. During his short stay at Lakehead University, he co-founded Genesis Genomics Inc. in 1999; a biotechnology company specialized in the analysis of mitochondrial genome and the design of biomarkers for the early detection of cancer. His main research interests include but not limited to Computational Intelligence (CI) and Machine Learning (ML) methods. He has a special interest in the applications of CI/ML methods to Computational Biology and Bioinformatics problems, such as: microarray analysis, protein analysis, oligonucleotide selection, bio-image analysis, and gene regulatory network analysis. He also has interests in applying CI/ML to such areas as: multiple valued logic algebras and circuits, wireless mobile networks, scheduling, grid-computing, and uncertainty modeling.