Bayesian Optimization Algorithm for the Non-unique Oligonucleotide Probe Selection Problem

Laleh Soltan Ghoraie, Robin Gras, Lili Wang, and Alioune Ngom

Bioinformatics and PRML Lab, Department of Computer Science,
University of Windsor,
401 Sunset Ave., Windsor, ON, N9B 3P4, Canada
{soltanl,rgras,wang111v,angom}@uwindsor.ca

Abstract. DNA microarrays are used in order to recognize the presence or absence of different biological components (*targets*) in a sample. Therefore, the design of the microarrays which includes selecting short Oligonucleotide sequences (*probes*) to be affixed on the surface of the microarray becomes a major issue. 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 presented approach considers integration of BOA and state-of-the-art heuristics introduced for the non-unique probe selection problem. This approach provides results that compare favorably with the state-of-the-art methods. It is also able to provide biologists with more information about the dependencies between the probe sequences of each dataset.

Keywords: Microarray, Probe Selection, Target, Estimation of Distribution Algorithm, Bayesian Optimization Algorithm, Heuristic.

1 Introduction

Microarrays are the tools typically used for measuring the expression levels of thousands of genes, in parallel. They are specifically applicable in performing many simultaneous gene expression experiments [10]. Gene expression level is measured based on the amount of mRNA sequences bound to their complementary sequences affixed on the surface of the microarray. This binding process is called *hybridization*. The complementary sequences are called *probes* which are typically short DNA strands about 8 to 30 bp [13]. Another important application of miocroarrays is the identification of unknown biological components in a sample [4]. Knowing the sequences affixed on the microarray and considering the hybridization pattern of sample, one can infer which target exists in the sample. These applications require finding a good *design* for microarrays. By microarray design, we mean finding the appropriate set of probes to be affixed on the surface

V. Kadirkamanathan et al. (Eds.): PRIB 2009, LNBI 5780, pp. 365–376, 2009.

of microarray. 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.

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 to which it hybridizes. It means that, in specified experimental conditions, the probe should not hybridize to other targets except for its intended target. However, finding unique probes are very difficult, especially for biological samples containing similar genetic sequences [4][5][6][8][10][11][12][13].

In the non-unique probe selection, each probe is considered to hybridize possibly to more than one target. Our focus in this paper is on the non-unique probe selection. We present a method to find the smallest possible set of probes capable of identifying the targets in a sample. It should be noticed that this minimal probe set is chosen regarding a target-probe incidence matrix consisting of candidate probes and the pattern of hybridization of targets to them. Computing the set of candidate probes (incidence matrix) among all the possible non-unique probes is not a trivial task [4]. Many parameters such as secondary structure, salt concentration, GC content, hybridization energy, and hybridization errors such as cross-hybridization, self-hybridization, and non-sensitive hybridization should be taken into account in computing the set of candidate probes for the oligonucleotide probe selection [12]. 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. The related work is reviewed in section 3. In section 4, we contribute our approach to solve non-unique probe selection problem. A review on the main concepts of Bayesian Optimization Algorithm (BOA) is also presented and its advantages over the Genetic Algorithms (GA) are discussed. Also, the heuristics which we have integrated into the BOA are discussed, and a new heuristic is presented. We discuss the results of our experiments in section 5. Finally, we conclude this research work with discussion of possible future research directions and open problems appears in section 6.

2 Problem Definition

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,...,t_3) and five probes (p_1,...,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 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

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

possible cardinalities) of 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. Moreover, the probe set $\{p_3, p_4, p_5\}$ can distinguish between all events in this case. A more formal definition of the probe selection problem is given below.

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) [5] [4]. A probe *separates* two subsets of targets if it hybridizes to either one of them. The probe selection is proven to be a NP-hard problem [2], and is considered as a variation of the combinatorial optimization problem *minimal set covering problem*.

The smallest incidence matrix in the literature contains about 256 targets and 2786 probes. The non-unique probe selection problem can be approached as an optimization problem. The objective function to be minimized is the number of probes (variables of the function), and the search space of the problem consists of $2^{numberofprobes}$ possible solutions which makes this problem very difficult to solve, even with powerful computers [8]. In this paper, we solve the single target case, and an EDA (Estimation Distribution Algorithms), named BOA (Bayesian Optimization Algorithm) integrated with some state-of-the-art probe selection heuristics, is used to design an efficient algorithm.

3 Previous Work

Several research works have been conducted in both unique and non-unique probe selection. Rash et al. [9] focused on the assumption of single targets in the sample. Considering the probes as substrings of original strings (genes), they used suffix tree method and Integer Linear Programming. Assuming the presence of multiple targets, Schliep et al. [10] introduced a fast heuristic which guaranteed the separation of up to a randomly chosen number N (e.g. $N = 500000$) of pairs of targets set. In this work, cross-hybridization and experimental errors were explicitly taken into account for the first time. Klau et al. [5] extended this work, and presented an ILP (Integer Linear Programming) formulation and a branch-and-cut algorithm to reduce the size of the chosen probe set.

The ILP formulation extended to a more general version which also includes the group separation [4]. Meneses et al. [6] used a two-phased heuristic to construct a solution and reduce its size for the case of single target. Ragle et al.

[8] 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. It does not use any *a priori* method to decrease the number of initial probes. Wang et al. [12] focused on the single target problem, and presented deterministic heuristics in order to solve the ILP formulation, and reduce the size of final probe set. They applied a model-based approach for coverage and separation in order to guide the search for the appropriate probe set in case of assuming single target in the sample. Recently, Wang et al. [11] presented a combination of the genetic algorithm and the selection functions used in [12], and obtained the results which are in some cases better than results of [8].

4 BOA and Non-unique Probe Selection

Our approach is based on the Bayesian Optimization Algorithm (BOA) in combination with a heuristic. Two of the heuristics, Dominated Row Covering (DRC) and Dominant Probe Selection (DPS), are the ones introduced in [12] for solving the non-unique probe selection problem. We also modify some of the function definitions of DRC, and introduce a new heuristic in order to capture more information.

4.1 Bayesian Optimization Algorithm

The BOA is an EDA (Estimation of Distribution Algorithm) method, first introduced by Pelikan [7]. EDAs are also called Probabilistic Model-Building Genetic Algorithms (PMBGA) which extend the concept of classical GAs. In the EDA optimization methods, the principle is to generate a sample of search space and use the information extracted from that sample to explore the search space more efficiently. The EDA approach is an iterative one consisting of these steps: (1) Initialization: a set of random solutions is generated (the first sample of search space); (2) Evaluation of the solutions quality; (3) Biased random choice of a subset of solutions such that higher quality solutions have more probability to be chosen; (4) Constructing a probabilistic model of the sample; (5) Use the model to generate a new set of solutions and go back to (2). In BOA, the constructed probabilistic model is a Bayesian Network. Considering a Bayesian Network as a Directed Acyclic Graph, the nodes represent the variables of the problem and the dependencies among the variables are simulated by the directed edges introduced to each node. Constructing a Bayesian Network allows discovering and representing the possible dependencies between the variables of the problem.

Some difficult optimization problems contain dependencies. Classical GAs has been shown not to be able to solve these category of problems [3]; But BOA approach has been more successful in solving them. It is interesting to apply BOA approach for the complex problem of non-unique probe selection optimization problem. In this problem each (binary)variable represents 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 [7] [1] 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 $O(2^k.n^{1.05})$ to guarantee the convergence.

There are several advantages in applying this new approach. First, BOA is known as an efficient way to solve the complex optimization problems. Therefore, it is interesting to compare it with other methods applied to the non-unique probe selection problem. Second, the 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 mechanism involved (See 5.2). 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.

4.2 Our Approach

In this section, we explain the details of our approach to solve the non-unique probe selection problem. Wang et al. [12] have introduced two heuristics in order to solve the non-unique probe selection problem. We integrated these heuristics into BOA in order to guarantee the *feasibility* of obtained solutions. A feasible solution is a solution which satisfies the constraints of coverage and separation of the non-unique probe selection defined in section 2. Since we discuss the case of single target in the sample, the separation constraint is applied on the target-pairs only. This means that we do not focus on the separation of all possible subsets of targets.

4.3 Heuristics

As mentioned above, our algorithm applies three heuristics in combination with the BOA. Two of the heuristics are those proposed by Wang et al. [12], namely, Dominated Row Covering (DRC), and Dominant Probe Selection (DPS). A third heuristic has also been used in our experiments, which we named *Sum of Dominated Row Covering*(*SDRC*). In this heuristic, we modified the definitions of the functions $C(p_j)$ (*coverage function*), and $S(p_j)$ (*separation function*) of DRC.

$$C(p_j) = \max_{t_i \in T_{p_j}} \{cov(p_j, t_i) \quad | \quad 1 \leq j \leq n\} \tag{1}$$

where Tp_j is the set of targets covered by p_j.

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

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

Before discussing our modifications, we describe the probe selection functions used in DRC (For further information on DPS selection functions, see Wang et al. [12]). Given the target-probe incidence matrix H, probe set $P = \{p_1,...,p_n\}$, and the target set $T=\{t_1,...,t_m\}$, the function *cov* and *sep* have been defined over $P \times T$ and $P \times T^2$, respectively, as following:

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

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

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

Function C favors the selection of probes that c_{min}-cover *dominated targets*. Target t_i dominates target t_j, if $P_{t_j} \subseteq P_{t_i}$. 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 $P_{t_{ij}} \subseteq P_{t_{kl}}$.

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 selection function $D(p_j)$ which has been defined as follows will indicate the degree of contribution of p_j.

$$D(p_j) = \max\{C(p_j), S(p_j)\} \quad | \quad 1 \leq j \leq n\} \tag{5}$$

The probes of highest value of $D(p_j)$ will be the candidate probes for the solution probe set. Calculation of the coverage and separation functions are given in Tables 2 and 3 based on DRC definitions in rows C and S, respectively [12]. We see, by definition of DRC functions, these four probes have the same score for the coverage of the dominated targets and the same score for the separation of the dominated target pairs, and $D(p_1) = D(p_3) = D(p_4) = D(p_5) = \frac{c_{min}}{3}$. Although, it can be noticed from 2 and 3 that each of these probes has a distinct covering and separating property. Therefore, these properties are not reflected by the definitions of current DRC functions. In order to capture this information, we modified the two functions of $C(p_j)$ and $S(p_j)$ to $C'(p_j)$ and $S'(p_j)$, respectively, in the *SDRC* (see Eq. 6 and 7 below). The values of $C'(p_j)$ and $S'(p_j)$ have also been calculated and presented in Tables 2 and 3. In the *SDRC*, the D score is calculated the same as D function in DRC (see Eq. 5).

Table 2. Coverage function table: C has been caculated based on the DRC definition, and C' based on the SDRC definition

	p_1	p_2	p_3	p_4	p_5	p_6
t_1	$\frac{c_{min}}{4}$	$\frac{c_{min}}{4}$	0	$\frac{c_{min}}{4}$	0	$\frac{c_{min}}{4}$
t_2	$\frac{c_{min}}{3}$	0	$\frac{c_{min}}{3}$	0	0	$\frac{c_{min}}{3}$
t_3	0	$\frac{c_{min}}{5}$	$\frac{c_{min}}{5}$	$\frac{c_{min}}{5}$	$\frac{c_{min}}{5}$	$\frac{c_{min}}{5}$
t_4	0	0	$\frac{c_{min}}{3}$	$\frac{c_{min}}{3}$	$\frac{c_{min}}{3}$	0
C	$\frac{c_{min}}{3}$	$\frac{c_{min}}{4}$	$\frac{c_{min}}{3}$	$\frac{c_{min}}{3}$	$\frac{c_{min}}{3}$	$\frac{c_{min}}{3}$
C'	$\frac{7c_{min}}{12}$	$\frac{9c_{min}}{20}$	$\frac{13c_{min}}{15}$	$\frac{47c_{min}}{60}$	$\frac{8c_{min}}{15}$	$\frac{47c_{min}}{60}$

Table 3. Separation function table: S has been calculated based on the DRC definition, and S' based on the SDRC definition

	p_1	p_2	p_3	p_4	p_5	p_6
t_{12}	0	$\frac{s_{min}}{3}$	$\frac{s_{min}}{3}$	$\frac{s_{min}}{3}$	0	0
t_{13}	$\frac{s_{min}}{3}$	0	$\frac{s_{min}}{3}$	0	$\frac{s_{min}}{3}$	0
t_{14}	$\frac{s_{min}}{5}$	$\frac{s_{min}}{5}$	$\frac{s_{min}}{5}$	0	$\frac{s_{min}}{5}$	$\frac{s_{min}}{5}$
t_{23}	$\frac{s_{min}}{4}$	$\frac{s_{min}}{4}$	0	$\frac{s_{min}}{4}$	$\frac{s_{min}}{4}$	0
t_{24}	$\frac{s_{min}}{4}$	0	0	$\frac{s_{min}}{4}$	$\frac{s_{min}}{4}$	$\frac{s_{min}}{4}$
t_{34}	0	$\frac{s_{min}}{2}$	0	0	0	$\frac{s_{min}}{2}$
S	$\frac{s_{min}}{3}$	$\frac{s_{min}}{2}$	$\frac{s_{min}}{3}$	$\frac{s_{min}}{3}$	$\frac{s_{min}}{3}$	$\frac{s_{min}}{2}$
S'	$\frac{31s_{min}}{30}$	$\frac{77s_{min}}{60}$	$\frac{13s_{min}}{15}$	$\frac{5s_{min}}{6}$	$\frac{31s_{min}}{30}$	$\frac{19s_{min}}{20}$

$$C'(p_j) = \sum_{t_i \in T_{p_j}} cov(p_j, t_i) \qquad 1 \leq j \leq n \tag{6}$$

$$S'(p_j) = \sum_{t_{ik} \in T^2_{p_j}} sep(p_j, t_{ik}) \qquad 1 \leq j \leq n \tag{7}$$

4.4 The Combination of BOA and Heuristics

We have applied the modified version of BOA to the non-unique probe selection problem. The goal is to find the minimum set of probe that satisfies the coverage and separation constraints. In each iterative step of BOA, we generate a population of solutions. Each solution is a representation of a set of probes, and is basically a string of zeros and ones. Each position in the string indicates a probe. The presence or absence of each probe in the solution is noted by 1 and 0, respectively. After generating the population, the feasibility of each solution is guaranteed by computing one of the heuristics described in section 4.3. That is, each solution in the current population is transformed in order to respect the problem constraints. All of the three applied heuristics include a *reduction* phase. Solutions are shortened in this phase, while maintaining their feasibility.

In order to measure the quality of the obtained solutions and distinguish the best and worst solutions in the population, an objective function should be defined. Since the goal is to find the minimal probe set in this problem, 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. The larger the objective function value, the higher the quality of the obtained solutions.

5 Results of Computational Experiments

We combined BOA and with heuristic DRC, DPS, and SDRC for non-unique probe selection problem. We noticed that we are able to improve the results obtained by the best methods in literature. It should be noticed that our approach

is more time-consuming than other approaches in the literature; But we did not focus on comparing our approach to the latest approaches from the aspect of the execution time, because the design of microarray is not a repetitive task. The main concern in this process is the quality of the design. Our programs were written in C++, and experiments were performed on Sharcnet systems [14].

5.1 Data Sets

The experiments were performed on ten artificial datasets named a1,..., a5, b1,..., b5, and two real datasets HIV1 and HIV2. These datasets have been used in experiments of all previous works mentioned in the section 3, except for the HIV1, and HIV2 that have not been used in [5][4]. The datasets and the related target-probe incidence matrices were kindly provided to us by Dr. Pardalos and Dr. Ragle [8]. Number of targets and probes of each data set are presented in Table 4. Along with this information, the number of virtual probes required for each dataset to guarantee the feasibility of the original probe set are included.

5.2 Results and Discussions

In all experiments, the parameters c_{min} and s_{min} were set to ten and five, respectively. Each run of BOA has been executed for 100 iterative steps. The number of probes in each dataset are the number of variables (n) used in the BOA. Based on the convergence condition of BOA, mentioned in the section 4.1, the population size should be of $O(2^k.n^{1.05})$. Two different series of experiments are performed, and the results are presented. In each series, we chose the population size for each dataset proportional to the number of the variables, which is sum of the number of real and the number of virtual probes of dataset. The considered level of dependency (k) among variables is simulated by a parameter named maximum incoming edges in the BOA software.

Experiments with the default parameters. First series of experiments have been performed with the default parameters of BOA [15]. For instance, the maximum number of incoming edges to each node was set to two, and the percentage of the offspring and parents in the population was set to 50. The results we obtain by applying this approach are presented in Table 4. The comparison between the results is based on the minimum set of probes obtained from each approach. We have named the combination of BOA and heuristics DRC, DPS, and SDRC respectively BOA+DRC, BOA+DPS, and BOA+SDRC. Three columns have been included related to experiments performed by state-of-the-art approaches Integer Linear Programming (ILP) [5][4], Optimal Cutting Plane Algorithm (OCP) [8], and Genetic Algorithm (DRC-GA) [11]. The last three columns show the improvement of our approach over each of the three latest approaches. The improvement is calculated by Eq. 8.

$$Imp = \frac{P_{min}^{BOA+DRC} - P_{min}^{Method}}{P_{min}^{Method}} \times 100 \tag{8}$$

where Method can be substituted by either ILP, OCP, or DRC-GA.

Table 4. 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 ($|V|$). The last three columns are showing the improvement of BOA+DRC over three methods ILP, OCP, and DRC-GA (see Eq. 8).

Set	$\|T\|$	$\|P\|$	$\|V\|$	ILP[5][4]	OCP[8]	DRC[11] -GA	BOA +SDRC	BOA +DPS	BOA +DRC	ILP	OCP	DRC -GA
a1	256	2786	6	503	509	502	503	503	502	-0.20	-1.37	0
a2	256	2821	2	519	494	490	492	491	490	-5.59	-0.81	0
a3	256	2871	16	516	543	534	535	533	533	+1.35	-2.02	-0.18
a4	256	2954	2	540	539	537	540	538	537	-0.55	-0.37	0
a5	256	2968	4	504	529	528	530	530	528	+4.76	-0.19	0
b1	400	6292	0	879	830	839	843	837	834	-5.12	+0.50	-0.60
b2	400	6283	1	938	842	852	853	849	846	-9.81	+0.47	-0.70
b3	400	6311	5	891	827	835	839	831	829	-6.96	+0.24	-0.72
b4	400	6223	0	915	873	879	877	877	875	-4.37	+0.23	-0.45
b5	400	6285	3	946	874	890	887	886	879	-7.08	+0.57	-1.23
HIV1	200	4806	20	-	451	450	452	450	450	-	-0.22	0
HIV2	200	4686	35	-	479	476	479	475	474	-	-1.04	-0.42

The calculated value of Imp is negative(positive) when BOA+DRC returns a probe set smaller(larger) than P_{min}^{Method}. Therefore, smaller value of Imp shows more efficiency of the BOA+DRC method. For instance, regarding Table 4 (last three columns), for dataset a3, our approach has obtained 0.18% and 2.02% better results (smaller probe set) than DRC-GA and OCP, respectively, and 1.35% worse result (larger probe set) than ILP.

As shown in the Table 4, the best results are obtained with the BOA+DRC, while we expected better results from the BOA+DPS, because the DPS has shown better performance on the non-unique probe selection [12]. The results obtained by the [8] are considered as the best ones in the literature for the non-unique probe selection problem. As shown in the 4, Wang et. al. [11] have recently reported the results (noted as DRC-GA) which are comparable to (and in most cases better than) [8].

Comparing our approach to all the three efficient approaches, we have been able to improve the result of non-unique probe selection for dataset HIV2, and obtain the shortest solution length of 474. The results we obtained for datasets a1, a2, a4, and HIV1 are also equal to the best results calculated for these datasets in the literature. Another comparison based on the number of datasets is presented in Table 5.

Another important advantage of our approach over other methods is that BOA can provide biologists with useful information about the dependencies between the probes of the dataset. In each experiment, we have stored the scheme of the relations between variables (probes) which have been found by BOA. As mentioned, by means of this information, we can realize which probes are related to each other. Therefore, we can conclude the targets, that these probes hybridize

Table 5. 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. In the column *average*, the average of improvements of our approach (illustrated in last three columns of Table 4) is presented.

	Worse	Equal	Better	Average
ILP	2	0	8	-3.36
OCP	5	0	7	-0.33
GA-DRC	0	5	7	-0.36

```
30 <-
31 <-
32 <- 1720, 4184
33 <- 3175, 3176
34 <-
35 <- 38, 7
36 <- 3, 90
37 <- 2822, 2819
38 <- 7, 4216
```

Fig. 1. Part of the BOA output for dataset HIV2: the discovered dependencies for probes 30 to 38 by BOA

to, also have correlations with each other. A part of these dependencies obtained for dataset HIV2 is presented in Figure 1. This Figure indicates parts of the output of the BOA software. Probes 30 to 38 and their dependencies to other probes are illustrated. As shown, no dependency has been discovered for probes 30, 31, and 34. Probe 32 has two incoming edges from probes 1720 and 4184. It means that when probes 1720 and 4184 are selected for the final probe set, probe 32 has high probability to also be selected for solving this problem.

Experiments for investigation of dependency. We conducted another series of experiments in order to study the effect of increasing the number of dependencies searched by BOA. The parameter *maximum incoming edges* represents this in BOA. As mentioned before, this parameter was set to two for previous experiments. We decided to increase this number to three and four, and repeat the experiments of BOA+DRC for some of the datasets. The results and the number of iterative steps to converge are shown in Table 6. We did not notice any improvements in results, but comparing cases of $k = 2$ and $k = 3$, the number of iterative steps to converge has been reduced. According to the results, it is possible that the obtained results are the global optimal solutions for some of the mentioned datasets. It is also possible that this problem does not contain high order dependencies. Therefore, search for higher order dependencies does not help to solve the problem. These should be further investigated with more experiments.

Table 6. Cardinality of minimal probe set for DRC+BOA: the experiment was repeated in order to investigate the effect of increasing the dependency parameter (k). By *gen* in the table, we mean the number of iterative steps of BOA to converge.

Set	$k = 2$	$k = 3$	$k = 4$
a1	502 gen:26	502 gen:17	502 gen:19
a2	490 gen:21	490 gen:20	490 gen:15
a3	533 gen:24	533 gen:19	533 gen:17
a4	537 gen:20	537 gen:17	537 gen:22
a5	528 gen:16	528 gen:13	528 gen:15

6 Conclusions (and Future Research)

In this paper, we presented a new approach for solving the non-unique probe selection problem. Our approach which is based on one of the EDAs named BOA obtains results that compare favorably with the state-of-the-art. Comparing to all the approaches deployed on the non-unique probe selection, our approach proved its efficiency. It obtained the smallest probe set for most datasets. Besides its high ability for optimization, our approach has another advantage over others which is its ability to indicate dependencies between the variables or probes for each dataset. This information can be of interest for biologists.

We also investigated the effect of increasing the dependencies between variables searched by BOA for some of the datasets. According to the presented results, it is possible that the results found for some of these datasets are the global optimal values. This requires more experiments and investigation. The non-unique probe selection has been discussed in this paper according the assumption of existence of single target in the sample. Therefore, one of the future works can be to focus on extending the problem with the assumption of multiple targets in the sample. Also, the discovered dependencies by our approach can be interpreted more precisely by biologists in order to detect more interesting information. As an extension to the presented work, we plan to incorporate several metrics into solution quality measure, and use a multi-objective optimization technique. One of the objectives can be the measure of ability of obtained solutions to recognize all targets present in the sample. This is referred to as *decoding* ability [10]. Using multi-objective optimization, parallelization techniques in the implementation can also be used in order to improve the running time of experiments considerably.

References

1. Goldberg, D.E.: The Design of Innovation: Lessons from and for Competent Genetic Algorithms. Kluwer Academic Publishers, Dordrecht (2002)
2. Garey, M., Johnson, D.: Computers and Intractability: A guide to the Theory of NP-Completeness. W. Freeman, San Francisco (1979)

3. Gras, R.: How Efficient Are Genetic Algorithms to Solve High Epistasis Deceptive Problems? In: Proc. 2008 IEEE Congress on Evolutionary Computation, Hong Kong, China, June 1-6, pp. 242–249 (2008)
4. Klau, G.W., Rahmann, S., Schliep, A., Vingron, M., Reinert, K.: Integer linear programming approaches for non-unique probe selection. Discrete Applied Mathematics 155, 840–856 (2007)
5. Klau, G.W., Rahmann, S., Schliep, A., Vingron, M., Reinert, K.: Optimal robust non-unique probe selection using integer linear programming. Bioinformatics 20, i186–i193 (2004)
6. Meneses, C.N., Pardalos, P.M., Ragle, M.A.: A new approach to the non-unique probe selection problem. Annals of Biomedical Engineering 35(4), 651–658 (2007)
7. Pelikan, M.: Bayesian Optimization Algorithm: From Single Level to Hierarchy. University of Illinois. PhD Thesis (2002)
8. Ragle, M.A., Smith, J.C., Pardalos, P.M.: An optimal cutting-plane algorithm for solving the non-unique probe selection problem. Annals of Biomedical Engineering 35(11), 2023–2030 (2007)
9. Rash, S., Gusfield, D.: String barcoding: uncovering optimal virus signatures. In: Annual Conference on Research in Computational Molecular Biology, pp. 254–261 (2002)
10. Schliep, A., Torney, D.C., Rahmann, S.: Group testing with DNA chips: generating designs and decoding experiments. In: Proc. IEEE Computer Society Bioinformatics Conference (CSB 2003), pp. 84–91 (2003)
11. Wang, L., Ngom, A., Gras, R.: Non-Unique Oligonucleotide Microarray Probe Selection Method Based on Genetic Algorithms. In: Proc. 2008 IEEE Congress on Evolutionary Computation, Hong Kong, China, June 1-6, pp. 1004–1010 (2008)
12. Wang, L., Ngom, A.: 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
13. Wang, L., Ngom, A., Gras, R.: Evolution strategy with greedy probe selection heuristics for the non-unique oligonucleotide probe selection problem. In: Proc. 2008 IEEE Symposiunm on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB 2008), pp. 54–61 (2008)
14. `http://www.sharcnet.ca/`
15. `http://www.cs.umsl.edu/~pelikan/software.html`