2009

Bayesian Optimization Algorithm for Non-unique Oligonucleotide Probe Selection

Laleh Soltan Ghoraie
University of Windsor

Bayesian Optimization Algorithm for Non-unique Oligonucleotide Probe Selection

by

Laleh Soltan Ghoraie

A Thesis
Submitted to the Faculty of Graduate Studies through the
School of Computer Science in Partial Fulfillment
of the Requirements for the Degree of Master of Science at the
University of Windsor

Windsor, Ontario, Canada
2009

Declaration of Co-Authorship/Previous Publications

I hereby declare that this thesis incorporates material that is result of joint research, as follows: *This thesis also incorporates the outcome of a joint research undertaken in collaboration with professor Dr. Alioune Ngom and Mrs. Lili Wang. The collaboration is covered in Chapters and of the thesis. In all cases, the key ideas, primary contributions, experimental designs, data analysis and interpretation, were performed by the author, and the contribution of co-authors was primaliry through the provision of corrections and constructive criticism.*

I am aware of the University of Windsor Senate Policy on Authorship and I certify that I have properly acknowledged the contribution of other researchers to my thesis,and have obtained written permission from each of the co-author(s) to include the above material(s) in my thesis.

I certify that, with the above qualification, this thesis, and the research to which it refers, is the product of my own work.

This thesis includes 2 original papers that have been previously published/submitted for publication in peer reviewed journals, as follows:

1. L. Soltan Ghoraie, R. Gras, L. Wang, A. Ngom, "Bayesian Optimization Algorithm

for the Non-unique Oligonucleotide Probe Selection Problem.", In Proceedings of the fourth IAPR International Conference on Pattern Recognition in Bioinformatics, Sheffield, UK, 365-376, 2009. (published)

2. L. Soltan Ghoraie, R. Gras, L. Wang, A. Ngom, "Optimal Decoding and Minimal Length for the Non-unique Oligonucleotide Probe Selection Problem.", 2009. (Submitted to the journal of Neurocomputing, special issue on Bioinformatics).

I hereby certify that I have obtained a written permission from the copyright owner(s) to include the above published material(s) in my thesis. I certify that the above material describes work completed during my registration as graduate student at the University of Windsor.

I certify that, to the best of my knowledge, my thesis does not infringe upon anyone's copyright nor violate any proprietary rights and that any ideas, techniques, quotations, or any other material from the work of other people included in my thesis, published or otherwise, are fully acknowledged in accordance with the standard referencing practices. Furthermore, to the extent that I have included copyrighted material that surpasses the bounds of fair dealing within the meaning of the Canada Copyright Act, I certify that I have obtained a written permission from the copyright owner(s) to include such material(s) in my thesis and have included copies of such copyright clearances to my appendix.

I declare that this is a true copy of my thesis, including any final revisions, as approved by my thesis committee and the Graduate Studies office, and that this thesis has not been submitted for a higher degree to any other University or Institution.

Abstract

One important application of DNA microarrays is measuring the expression levels of genes. 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 thesis 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* named *Bayesian Optimization Algorithm (BOA)* to this problem. This approach considers integration of BOA and one simple heuristic. The obtained results compare favorably with the state-of-the-art methods. We also present application of our method in integration with *decoding* approach in a multiobjective optimization framework for solving the problem in the case of multiple targets in the sample.

Dedication

I would like to dedicate this thesis to my wonderful parents.

Acknowledgements

First I would like to thank my family for their endless support, compassion, and kindness.

I am very grateful to my supervisor, Dr. Robin Gras for his valuable guidance during my thesis work and through completion of my degree requirements and research. Without his support, this thesis could not have been completed. I also would like to thank members of my master committee, Dr. Alioune Ngom, School of Computer Science, Dr. Lisa Porter, Department of Biological Sciences, for being in my committee, and their constructive criticism, helpful advices, and guidance, and to Dr. Luis Rueda, the chair of the committee.

I acknowledge the financial support from my supervisor, Dr. Robin Gras, in the form of research assistantship through NSERC; from the School of Computer Science, in the form of graduate assistantship.

I gratefully acknowledge assistance of Dr. Alexander Schliep, Department of Computer Science and BioMaPS Institute for Quantitative Biology, Rutgers University, and Mr. Ole Schulz-Trieglaff, PhD candidate in Free University of Berlin, for providing me by their group's software, TCPD (a modified version of MCPD [56]), and their helpful comments, that resulted in improvement of this investigation.

I am also thankful of Dr. Panos Pardalos and Dr. Michelle A. Ragle, Department of Industrial and System Engineering, University of Florida, who kindly provided us the data required for the experiments of the thesis.

Contents

List of Figures

List of Tables

List of Algorithms

Chapter 1

Introduction

Microarrays are tools used for performing many hybridization experiments in parallel. As noted by [43], two main applications are considered for microarrays. 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 to 30 bp [53]. The second important application of miocroarrays is the identification of unknown biological components in a sample [21]. 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 [43].

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. The probe selection problem is discussed in this thesis. Before addressing the problem, we present a general introduction of the microarrays.

1.1 Functional Genomics and Microarrays

The research field of functional genomics aims to understand functions of genes, their interactions, and how they are regulated [38]. Experiments are conducted in this research field in order to obtain knowledge about state of a genomic system. One of the major related tasks is to obtain knowledge about gene expression and regulation. Several techniques are applied in order to measure the gene expression. These techniques mostly focus on the quantification of mRNA molecules in the cell [38]. All of these techniques are based on the fact that the nucleic acids hybridize to their complements. The constructed hybrid in these hybridization experiments refer to the built double-stranded molecule from one DNA strand and one RNA strand.

Some of the frequently used techniques of gene expression measurement are "northern blot analysis", "RPA (ribonuclease protection assay)", "RT-PCR (reverse transcription-polymerase chain reaction)", "SAGE (serial analysis of gene expression)", and "in-situ hybridization". For further information, refer to [48] and [29].

The disadvantage of these techniques is that they focus on a single gene or a few genes at a time. On the other hand, the scope of the investigations was extended from a single gene to studying all the genes at once. Therefore, the current techniques were not able to satisfy new requirements. This caused the development of "DNA microarrays", which became a proper research tool for functional genomics.

DNA Microarrays or DNA chips (Figure 1.1) are arrays of many DNA molecules (probes) on a quartz, glass, or nylon surface [38]. The probes are segments of known genes affixed in the locations called spots on the chip. Targets are mRNA extracted molecules from a blood sample or tissue, which are labeled with a fluorescent or radioactive dye. The function of microarrays are based on the construction of RNA-DNA-hybrids or double-stranded DNA.

In the general process of microarray experiments, the targets are allowed to hybridize to the probes of the chip. In case of finding their complementary sequences in a sample of targets, probes hybridize to the targets. The targets which have not hybridized are washed away from the chip, and the amount of hybridization in each spot which can be recognized by the intensity of the fluorescence or radioactivity, can be used in order to measure the

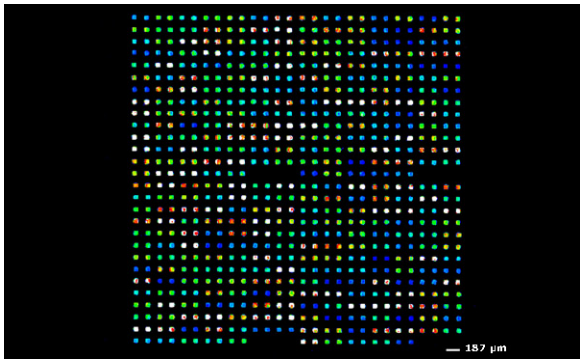

Figure 1.1: DNA Microarray

gene expression level. The details of this experiment depends on the microarrays platform.

Four main technology platforms of microarrays that are listed by [38] as follows: "nylon membrane arrays" or "radioactive filters", "cDNA arrays" or "red/green arrays", "polynucleotide", and "oligonucleotide arrays". The last one is the subject of this thesis discussion.

[38] distinguishes three important applications of oligonucleotide arrays or DNA chips: 1-gene expression measurements, which is the most important application, and the DNA chips are commonly used for these measurements because they have high number of spots on a small surface which allow conducting many expression experiments in parallel. The researchers are able to apply the obtained information from these experiments to study gene functions [20], to discover genetic causes of diseases [13], etc. 2-sequencing by hybridization (SBH) [26], and 3-determination of single nucleotide polymorphisms (SNP)s [5] [23] [16].

An oligonucleotide chip follows these steps from production to analysis of obtained hybridization data: "chip production", "target preparation", "hybridization, washing, and staining", "data acquisition and analysis".

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. 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. On the other hand, smaller set of probes designed for the microarrays leads to more cost-efficient experiments, because the number of probes are

proportional to the number of hybridizations that are performed.

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 that 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 [21] [22] [30] [35] [43] [51] [52] [53]. When many targets are similar, experimental errors increase. In these cases, 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 [53]. Minimizing the probe set is an important objective. Smaller microarray designs occupy less space on the surface of microarray. This leads to use smaller chips, and reduce the costs considerably [43].

Our focus in this thesis 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, our task is to analyze and minimize the design in order to select the best possible probe set. The initially given 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 [21]. Computing the initial target-probe incidence matrix in not a trivial task [22].

1.2 Probe Design

As mentioned, we are given a probe design including *good* probes which are candidates for probe selection. The design is given as target-probe incidence matrices, and our task in to minimize it. Computing a proper design is not a trivial task. Although this thesis does not include the discussion on the computing of the incidence matrices, in this section, we briefly explain the important parameters which are considered for this design computation.

Among a lot of possible non-unique probes, some are selected as the candidate probes for the chip design. These probes should satisfy criteria of: *Homogeneity*, *Sensitivity*, and *Specificity* [12] [30].

Homogeneity property discusses about the melting temperature of the probes [12] [30] denoted by T_m. T_m is a temperature at which half of the DNA molecules have been separated to single-stranded molecules [42]. In order to satisfy homogeneity property, the melting temperature of all the candidate probes should be within a predefined range close to the experiment temperature. This is for ensuring that the probes will hybridize to the intended targets at temperature about the experiment temperature. This property causes uniform performance among the probes [49].

In other words, the candidate probes should be isothermal, that is, they should behave similarly under conditions of hybridization experiment, such as, salt concentration and temperature [47]. The melting temperature is affected by different factors such as salt concentration of the solution and the base composition of the DNA. Also, DNA containing many G-C pairs has a higher melting temperature than one with more A-T pairs [31] [32].

A *sensitive* probe returns a strong signal when it is beside its complementary target sequence in the sample. Some probes are self-complementary which means they can fold back on themselves and this decreases the sensitivity of the probe [47] [49]. A probe prone to self-complementarity is demonstrated in the Figure 1.2 [12]. The self-complementary probes form secondary structures [30]. Gibbs free energy is a measure that is applied in order to predict the stability of secondary structure. The nearest neighbor model can be used to compute free energy.

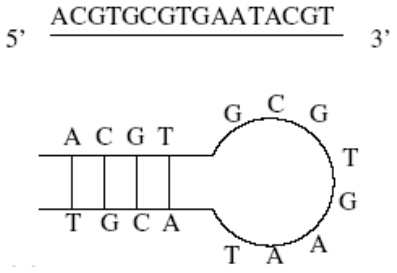

Figure 1.2: A probe prone to self-complementarity

Specific probes are unique ones for each gene of a genome. High specificity decreases cross-hybridization. Cross-hybridization happens when probes hybridize to targets other than their specified ones. The probes containing repetitive sequences are likely to cross-hybridize [30]. In order to increase the specificity of the candidate probe set, probes that contain repetitive sequences are filtered. This can be performed by means of softwares such as RepeatMasker [57] which detects the repetitive sequences [47].

A common similarity measure based on *Hamming Distance* is used for identifying specific probes. For two strings of a and b, the Hamming Distance ($H(a, b)$) is defined as the number of corresponding positions in which two strings have different characters. For instance, if $a = 10111101$ and $b = 11111000$, then $H(a, b) = 3$. The process of specificity check is computationally expensive [12]. The brute force approach considers all the genomes of length n, and searches for all the probes of length m in order to ensure the Hamming Distances are large enough. Time for this process is of $O(mn^2)$.

Except for these three major factors, there are three other constraints proposed by [28] which can be considered to improve the quality of the candidate probe set [12]: (1) probes should not contain any of the single bases (A, T, C or G) for more than 50% of their size; (2) probes should not contain contiguous sequence of As and Ts or Cs and Gs in regions of more than 25% of the probe size; (3) GC-content should be between 40% and 60% of the probe sequence.

As mentioned, many parameters such as secondary structure, salt concentration, GC content, hybridization energy, etc. influence the quality of the probes hybridization [43], 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 [22]. 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 [52]. Candidate probes also should neither be self-complementary nor should cross-hybridize [22].

1.3 Non-unique Oligonucleotide Probe Selection

As mentioned before, a unique probe hybridizes to only one target. Due to the difficulty in finding unique probes for closely related gene families, the unique probe selection approach is impractical for many datasets. The alternative is non-unique probe selection, in which a probe can hybridize to more than one target. Our focus in this work in the non-unique probe selection problem.

A formal definition of the non-unique 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*) [22] [21]. 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 [43]. 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 $p \in P(S)\Delta P(T)$ [43]. 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 $|P(S)\Delta P(T)| \geq s_{min}$.

This problem can be considered in two cases of single and multiple targets in the sample. We illustrate the two cases of 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 1.1). The incidence matrix contains the "good probes" and their hybridization pattern to targets. the good probes are selected in an earlier step named probe design (explained in section 1.2) .

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.

Table 1.1: 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

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. Moreover, 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 7.1.

The probe selection is proven to be a NP-hard problem [11], 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 of single target, although multiple targets in the sample is 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.

1.4 Contribution of this thesis

As mentioned, the non-unique probe selection 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 impossible to solve exhaustively, even with

powerful computers [35]. We propose a method based on an EDA (Estimation Distribution Algorithms), named BOA (Bayesian Optimization Algorithm)(see section 3.2) integrated with simple probe selection heuristics 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.

The heuristics used in integration with the BOA are Dominated Row Covering (DRC) and Dominant Probe Selection (DPS) which were proposed in [52] for solving the problem of non-unique probe selection (see section 4.2). Also, we propose a new heuristic named Sum of Dominated Row Covering (SDRC), and apply it for a series of experiments. The non-unique probe selection problem has been considered as optimization problems 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 of this problem is minimizing the number of selected probes. The results of our experiments compare favorably with the state-of-the-art methods.

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 for solving multiobjective optimization problems efficiently by means of evolutionary-based algorithms such as BOA (see section 5). 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 [43] (see section 6).

1.5 Organization of this thesis

This thesis is organized as follows. Chapter 2 provides a detailed review on the non-unique probe selection problem. Then, in subsequent chapters 3.2, 4.2, 6, and 5, we describe the fundamentals of our algorithm. A review on the main concepts of Bayesian Optimization Algorithm (BOA) is presented in chapter 3.2, and its advantages over the Genetic Algorithms (GA) are discussed. Also, the heuristics which we have integrated into the BOA are discussed in chapter 4.2. At the end of this section, we explain how and why we integrate these heuristics into the BOA in section 4.2. The multiobjective optimization technique and decoding idea applied in this work are discussed in chapters 5 and 6, respectively. We discuss the results of our experiments in chapter 7.3.2. Finally, we conclude this research work with discussion of possible future research directions and open problems appears in the sections 8.1 and 8.2 of the chapter 8.2, respectively.

Chapter 2

Review of Literature

The probes for hybridization experiments were selected mostly randomly or based on the frequency of occurance of probes sequences in the genes before the work of Herweig et al. [18]. Other criteria such as G+C content [4] [9], and free energy and melting temperature [27] were also considered. Herwig et al. [18] emphasized on the importance of selecting good and informative probes for the experiments, and formulating the problem of probe selection to be studied systematically. Their work which was focused on the unique probe selection, was the first one that considered the problem as an explicit optimization problem. They presented an information-theoretical approach based on entropy maximization to this problem. Their simple greedy heuristic based on clustering and entropy achieved probe sets of higher quality than the sets chosen randomly or based on frequency.

Borneman et al. [2] introduced two alternative formulations of the non-unique probe selection problem, called Minimum Cost Probe Set (MCPS), and other called Maximum Distinguishing Probe Set (MDPS). The first one focuses on finding a minimal probe set that is able to identify all the given clones. In the second one, for a given k, we focus on finding a probe set that is able to distinguish between maximum number of clone pairs. These two problems are NP-hard problems and variants of the *set cover* problem [19]. Borneman et al. [2] proposed two heuristic algorithms for solving these problem. The proposed heuristic

for MDPS is based on simulated annealing, and the one for MCPS is based on Lagrangian relaxation.

The work of Rash and Gusfield [40] was based on string barcoding problem which is useful in identifying an unknown string as one of a set of known strings. Rash and Gusfield considered genes as strings and the probes as substrings of these original strings. They used suffix tree to identify the critical substrings and eventually to reduce the number of variables in an Integer Linear Programming formulation. The ILP formulation was used to solve the optimization problem. They applied CPLEX [58] to solve the ILP problem.

The works [40] and [2] addressed the application of non-unique probes; But they did not consider the errors, and also simplified the problem by assuming at most one target to be present in the sample. As mentioned, this case is a simplified version of the non-unique probe selection problem.

Schliep et al. [43] proposed a three-stepped methodology (Figure 2.1 [43]) for microarray design based on a group testing approach [8]. They explained that the selection of candidate probes was based on an extended version of the longest common factor method [36]. The most suitable design candidates were selected and the target-probe incidence matrix was computed. They introduced a fast heuristic in order to select a minimal probe set that was able to distinguish between most targets sets of small cardinality. 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. $N = 500000$) 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 the sample. In this work, cross-hybridization and experimental errors were explicitly taken into account for the first time.

Klau et al. [22] stated the ILP formulation for the non-unique probe selection problem, and used a branch-and-cut algorithm formulation for solving the group separation problem which guaranteed separation of all possible target sets. However, the preliminary implementation of Klau et al. was capable of separating only the target pairs. Compared to the

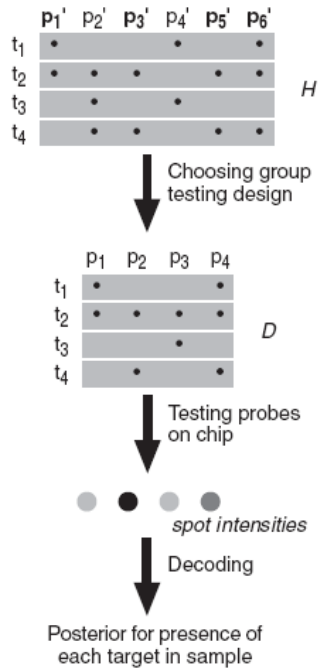

Figure 2.1: An overview of the three-stepped methodology proposed by [43]

heuristic proposed by Schliep et al. [43], this approach resulted in considerable reduction in the cardinality of the final probe set. CPLEX [58] was applied to solve the ILP. They also measured the decoding ability of the obtained probe set, and noticed a mild reduction. Years later, klau et al. also solved the more general version of ILP formulation which also includes all the group separation constraints [21] in which groups correspond to multiple targets. By this extension, the assumption of the multiple targets was realized.

As mentioned, the ILP formulation for the non-unique probe selection was first proposed in [22] and [21]. Target-probe incidenc matrix (H_{ij}), n probes and m targets and the constraints c_{min} and s_{min} are given. A set of binary variables (x_j) are considered corresponding to the probes (p_j) where $1 \leq j \leq n$, and n is number of probes. $x_j = 1$ if the probe p_j is present, otherwise $x_j = 0$. The non-unique probe selection problem is formulated as follows:

$$Minimize : \sum_{j=1}^{n} x_j \tag{2.1}$$

subject to:

$$x_j \in \{0,1\} \quad 1 \leq j \leq n \tag{2.2}$$

$$\sum_{j=1}^{n} H_{ij}xj \geq c_{min} \quad 1 \leq i \leq m \tag{2.3}$$

$$\sum_{j=1}^{n} |H_{ij} - H_{kj}|x_j \geq s_{min} \quad 1 \leq i < k \leq m \tag{2.4}$$

Function 2.1 indicates the minimization problem of probes. The constraint 2.2 restricts the variables of the problem to binary-valued ones. The coverage and separation constraints are demonstrated by 2.3 and 2.4, respectively. In this version of ILP, the constraint 2.4 indicates the single target case of the problem. There is another version of ILP formulation, proposed in the [21], which contains group separation constraints. Therefore, it covers the case of multiple targets in the sample. This formulation is as follows:

$$Minimize : \sum_{j=1}^{n} x_j. \tag{2.5}$$

subject to:

$$x_j \in \{0,1\} \quad 1 \leq j \leq n \tag{2.6}$$

$$\sum_{j=1}^{n} |\omega_j^{t_x^a} - \omega_j^{t_y^a}|x_j \geq \min \left\{ d, \sum_{j=1}^{n} |\omega_j^{t_x^a} - \omega_j^{t_y^a}| \right\} \quad \forall (t_x^a, t_y^a) \in \{2^T \times 2^T\}, \tag{2.7}$$

$$|t_x^a|, |t_y^a| \leq d_{max}, \quad t_x^a \neq t_y^a \tag{2.8}$$

where $c_{min} = s_{min} = d$. t_x^a and t_y^a are two sets of targets. The constraints of 2.7 are the group separation constraints (multiple targets case) and also cover the single target case. If $t_x^a = \emptyset$ and $t_y^a = t_i$ for $1 \leq i \leq m$, the Equation 2.7 and 2.8 satisfies the coverage constraint. Here, the ILP aims to select two sets of targets which cause the maximal violated group inequality [21].

Gasieniec et al. [12] proposed a new direction to confront with the probe selection problem. They introduced an efficient algorithm named RANDPS that randomly selects a small number of probes. They conducted their experiments by either one probe (unique probe) or multiple probes (non-unique probes), and the assumption of only one target in the sample. Their algorithm takes a set of known genes as input and instead of checking all possible probes, it randomly picks probes based on some minimal criteria checking. Considering non-exhaustive search resulted in a algorithm which is efficient in time and space. The experiments results show that a single probe is sufficient for identifying almost all the genes uniquely, and the others need at most two probes [12].

Wang et al. [52] has discussed the non-unique probe selection theoretically. The problem was presented as selection of a d-disjunct submatrix from the original (binary) target-probe incidence matrix. The submatrix should include the same number of rows (targets) and minimum possible number of columns (probes). Wang et. al showed that this minimization problem is MAX SNP-complete, but has a polynomial-time approximation with the performance ratio of $1 + \frac{2}{(d+1)}$. This minimization is polynomial-time solvable when every probe hybridizes to exactly two targets.

In [6], Deng et al. considered the group testing approach in studying the non-unique probe selection problem. The non-unique probe selection problem is presented in three steps in this work: 1) Collection of a large set of non-unique probes. 2) Find a minimum subset of probes to identify up to d viruses. 3) Testing the decoding ability of the result probe set. Deng et. al focused on the minimization problem (step 2). [6] was a survey of the computational complexity of the problem considering a non-adaptive group testing design approach. According to this survey, the best-known group testing design has been within a factor of $O(\log d)$ from the lower bound and the best-known approximation for the non-unique probe selection in within a factor of $O(\log n)$ from optimal solution [6].

Thai et al. [50] focused on the DNA library screening which requires efficient pooling designs in order to be able to recognize the positive and negative clones. The design of the decoding algorithm to determine whether a clone is positive regarding the design is a challenging task. The challenge is due to the experimental errors and presence of inhibitors

that are clones that neutralize positive clones. The novel decoding algorithm which is proposed in this work identifies all the positives in the presence of errors and inhibitors.

Deng et al. [7] proposed algorithms for the non-unique probe selection based on the Integer Linear Programming. Their focus was on the minimization problem while they paid attention to the decoding ability of the obtained solution. Their design algorithm first builds a matrix close to a d-disjunct matrix matrix using one ILP. Then another ILP is formulated for finding the violations of d-disjunctness. Addressing the violations and formulating another ILP for finding more violations is done recursively until all the violations are resolved. They claimed that their decoding algorithm is able to identify up to d targets in the sample with at most k experimental errors, and the algorithm complexity is $O(hn)$ in which h is the number of selected probes. They also claimed that their decoding algorithm is much faster than the other methods using d-separable matrix.

Focusing on the single target case, Meneses et al. [30] 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. In the first phase, a feasible solution is constructed for the ILP formula presented in [22]. Then iteratively this solution is reduced by removing probes while the solution still satisfies the coverage and separation constraints. This algorithm outperformed the method of [22] for the largest experimented dataset; But for the smaller datasets the obtained solutions included more probes than results in [22].

Ragle et al. [35] also based the work on the ILP formula presented in [22] and 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. Their method was able to reduce the cardinality of the final probe set by 20% compared to the state-of-art methods.

Wang et al. [52] 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. [51] presented a combination of the genetic algorithm and the selection functions used in [52], and obtained results which are in most cases better than results of [35].

Chapter 3

Estimation of Distribution Algorithm named Bayesian Optimization Algorithm

Genetic algorithms are optimization methods based on the selection and recombination operators. The partial solutions of a problem, called *building blocks*, are manipulated by the selection and recombination methods. These two mechanisms rebuild and mix the building blocks [34]. The general and fixed recombination operators often cause breaking the building blocks and loosing important information. This can lead to convergence to a local optimum. The problem of building block disruption is named *linkage problem*.

The linkage problem became an important deficiency of classic genetic algorithms. This deficiency caused the classic genetic algorithms not to be able to solve even problems composed of simple partial subproblems [34]. Mainly two classes methods proposed to prevent the linkage problem and disruption of partial solutions of the problem. First class of methods are focused on changing the representation of solutions or modifying the recombination operators. The second class are focused on finding ways to extract information from the promising data samples and use the information to generate new solutions. EDA was an approach which proposed in order to resolve the deficiency of the classic genetic algorithms, and was categorized as a technique of the second class.

3.1 Estimation of Distribution Algorithm

EDA (Estimation of Distribution Algorithm) method was introduced by Mühlenbein and Paaß [33] [24]. EDAs are also called Probabilistic Model-Building Genetic Algorithms (PMBGA) 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; 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

3.2 Bayesian Optimization Algorithm

In BOA, which was first proposed by Pelikan [34], 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. Figure 3.1 [59] displays the main iteration of the BOA.

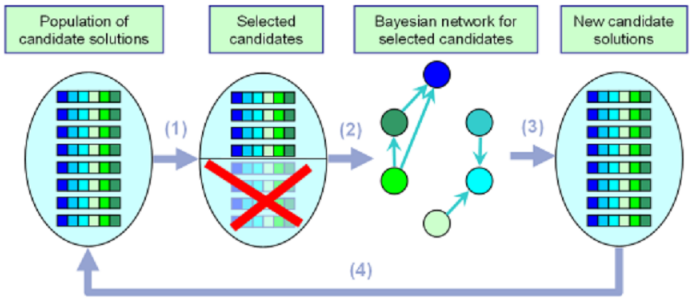

Figure 3.1: The main iteration of BOA

Based on the generic algorithm of EDA, in BOA, a probabilistic model which is a Bayesian Network is constructed in step (5) (See Algorithm 1). 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 [17]. After constructing the network, the joint probabilities of the variables 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 [14]. 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 non-unique 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 [34] [10] have proven that when the number of variables and the maximum number of dependencies for any variable are n and k, respectively, the size of the sample should be about of $O(2^k.n^{1.05})$ to guarantee convergence with a given probability.

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.

Chapter 4

Heuristics

4.1 Introduction

Our algorithm applies three heuristics in combination with the BOA. Two of the heuristics are those proposed by Wang et al. [52], 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. As mentioned above, our algorithm integrates simple heuristics to the BOA.

4.1.1 Dominated Row Covering Heuristic

The heuristic Dominated Row Covering (DRC) was proposed by Wang et al. [52]. Given the target-probe incidence matrix H, probe set $P = \{p_1,...,p_n\}$, and the target set $T=\{t_1,...,t_m\}$, two main functions $C(p_j)$ (*coverage function*) and $S(p_j)$ (*separation function*) have been defined for this heuristic as follows.

$$C(p_j) = \max_{t_i \in T_{p_j}} \{cov(p_j, t_i) \quad | \quad 1 \leq j \leq n\} \tag{4.1}$$

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

$$S(p_j) = \max_{t_{ik} \in T^2_{p_j}} \{sep(p_j, t_{ik}) \quad | \quad 1 \leq j \leq n\} \tag{4.2}$$

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 $P_{t_j} \subseteq P_{t_i}$. 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 functions cov and sep, respectively.

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

$$cov(p_j, t_i) = h_{ij} \times \frac{c_{min}}{|P_{t_i}|}, \qquad p_j \in P_{t_i}, t_i \in T \tag{4.3}$$

$$sep(p_j, t_{ik}) = |h_{ij} - h_{kj}| \times \frac{s_{min}}{|P_{t_{ik}}|}, \qquad p_j \in P_{t_{ik}}, t_{ik} \in T^2 \tag{4.4}$$

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}.

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.

$$D(p_j) = \max\{C(p_j), S(p_j)\} \quad | \quad 1 \leq j \leq n\} \tag{4.5}$$

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 4.1, in Tables 4.2 and 4.3, respectively [52].

Table 4.1: 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 4.2: 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	$\frac{c_{min}}{4}$	$\frac{c_{min}}{4}$	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	$\frac{c_{min}}{3}$	$\frac{c_{min}}{4}$	$\frac{c_{min}}{3}$	$\frac{c_{min}}{3}$	$\frac{c_{min}}{3}$	$\frac{c_{min}}{3}$

Table 4.3: 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	$\frac{s_{min}}{3}$	$\frac{s_{min}}{3}$	$\frac{s_{min}}{3}$	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	$\frac{s_{min}}{3}$	$\frac{s_{min}}{2}$	$\frac{s_{min}}{3}$	$\frac{s_{min}}{3}$	$\frac{s_{min}}{3}$	$\frac{s_{min}}{2}$

The DRC algorithm consists of three phases of : *Initialization*, *Construction*, and *Reduc-*

tion. 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 (see 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 (see 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 hope to obtain a minimal feasible solution (P_{red}).

4.1.2 Sum of Dominated Row Covering Heuristic

According to DRC algorithm (section 4.1.1), 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 were given in Tables 4.2 and 4.3 based on DRC definitions in rows C and S, respectively [52]. We see, by definition of DRC functions, the four probes of p_1, p_3, p_4, and p_5 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 4.2 and 4.3 that each of these probes has a distinct covering and separating property. These properties are not reflected by the definitions of current DRC functions.

In order to capture this information, we modified and redefined the two functions of $C(p_j)$ and $S(p_j)$, in the *SDRC* (see Equation 4.6 and 7.2 below). The values of $C(p_j)$ and $S(p_j)$ have been recalculated and presented in Tables 4.4 and 4.5. In the *SDRC*, the D score is calculated the same as D function in DRC (see Equation 6.5).

$$C(p_j) = \sum_{t_i \in T_{p_j}} cov(p_j, t_i) \qquad 1 \leq j \leq n \tag{4.6}$$

$$S(p_j) = \sum_{t_{ik} \in T^2_{p_j}} sep(p_j, t_{ik}) \qquad 1 \leq j \leq n \tag{4.7}$$

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 Equations (6.3)–(6.5)

$P_{\text{ini}} \leftarrow \{p \in P \mid D(p) = 1\}$ */* *essential probes only* */*

/* Construction Phase */

$P_{\text{sol}} \leftarrow P_{\text{ini}}$

Sort $P \setminus P_{\text{sol}}$ 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

$P_{\text{sol}} \leftarrow P_{\text{sol}} \cup \bigcup_{l=1}^{l=n_i} \{\text{next highest-degree probe } p_l \in P \setminus P_{\text{sol}} \text{ that covers } t_i\}$

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}

$P_{\text{sol}} \leftarrow P_{\text{sol}} \cup \bigcup_{l=1}^{l=n_{ik}} \{\text{next highest-degree probe } p_l \in P \setminus P_{\text{sol}} \text{ that separates } t_{ik}\}$

end

/* Reduction Phase */

$P_{\min} \leftarrow P_{\text{sol}}$

$H \leftarrow H|_{P_{\min}}$, */* *restriction of* H *to probes in* $P_{\min}$ */*

Compute $D(p)$ for all $p \in P_{\min}$

Sort $P_{\text{del}} \leftarrow \{p \in P_{\min} \mid D(p) < 1\}$ in increasing $D(p)$

if $P_{\min} \setminus \{p\}$ *is feasible for each* $p \in P_{\text{del}}$ **then**

$P_{\min} \leftarrow P_{\min} \setminus \{p\}$

end

Return final $P_{\min}$

end

4.1.3 Dominant Probe Selection Heuristic

The heuristic Dominant Probe Selection (DPS), proposed by Wang et al. [52], favors the selection of *dominant probes*. p_j *dominates* p_l if $T_{p_l} \subset T_{p_j}$. As it was shown in the Table

Table 4.4: Coverage function table: C has been calculated 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	$\frac{7c_{min}}{12}$	$\frac{9c_{min}}{20}$	$\frac{13c_{min}}{15}$	$\frac{47c_{min}}{60}$	$\frac{8c_{min}}{15}$	$\frac{47c_{min}}{60}$

Table 4.5: Separation function table: S has been calculated 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	$\frac{31s_{min}}{30}$	$\frac{77s_{min}}{60}$	$\frac{13s_{min}}{15}$	$\frac{5s_{min}}{6}$	$\frac{31s_{min}}{30}$	$\frac{19s_{min}}{20}$

4.1, $T_{p_1} = \{t_1, t_2\}$ and $T_{P_6} = \{t_1, t_2, t_3\}$. Therefore, $T_{p_1} \subset T_{p_6}$, and p_6 dominates p_1.

By selecting dominant probes instead of dominated probes, more targets can be covered. To favor the selection of a dominant probe that has the same degree as some of its dominated probes, the definitions of functions cov and sep (Equations 6.3 and 6.4) have been modified in order to give higher value to dominant probe p rather than the dominated probes. This is possible with penalizing each entry in Tables 4.2 and 4.3 by an amount that takes into account the number of targets covered and the number of target-pairs separated by a given probe. The new cov and sep functions are respectively as follows

$$\mathrm{cov}(p_j, t_i) = h_{ij} \times \frac{c_{\min}}{|P_{t_i}|} \times \frac{1}{m - |T_{p_j}|} \ , \tag{4.8}$$

where $p_j \in P_{t_i}$, $t_i \in T$, and $\mathrm{cov}(p_j, t_i) \in [0, 1]$; P_{t_i} is the set of probes hybridizing to target

t_i, T_{p_j} in the penalty term is the set of targets covered by p_j, and m is the number of targets. By new definition of cov function, probes that cover fewer targets are penalized more than probes that cover more targets.

$$\text{sep}(p_j, t_{ik}) = |h_{ij} - h_{kj}| \times \frac{s_{\min}}{|P_{t_{ik}}|} \times \frac{1}{\frac{m(m-1)}{2} - |T^2_{p_j}|} \quad , \tag{4.9}$$

where $p_j \in P_{t_{ik}}$, $t_{ik} \in T^2$, and $\text{sep}(p_j, t_{ik}) \in [0, 1]$; $P_{t_{ik}}$ is the set of probes separating target-pair t_{ik} and $T^2_{p_j}$ is the set of target-pairs separated by p_j. By new definition of sep function, probes that separate fewer target-pairs are penalized more than probes that separate more target-pairs.

The difference between DPS and DRC heuristics is in the definitions of functions cov and sep, as described above.

4.2 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 one of the heuristics described above. A minimum set of probes should satisfy the coverage and separation constraints. Since the probe set found by BOA does not guarantee the constraints satisfaction, we have applied the heuristics in order to guarantee this issue.

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 6).

This results in forming a one-objective and a two-objective optimization problem in cases of single targets and multiple targets in the sample respectively. On the other hand, appropriate multiobjective optimization technique (see chapter 5) should be applied to solve the two-objective problem.

Chapter 5

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 fitness 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 fitness 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.

A classical approach to deal with this issue 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. The property is range-independence. In most of the complex multiobjective problems, each objective has an effective range, and the function ranges is non-commensurable [44]. 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) which is demonstrated in Algorithm 3. In this method, the fitness values of the solutions for each objective (noted by "O" in the Algorithm 3) are extracted and listed separately. In the Algorithm 3, the lists are noted as "FINESS-LIST"s. The lists are sorted, and based on the position (noted by "P" in Algorithm 3) of each fitness value, a rank (noted by "SOLUTION.RANK" in the Algorithm 3) is assigned to the fitness value of the solution. For each solution, different ranks obtained by sorting each list of objectives is averaged (indicated by ** in the Algorithm 3). Since each objective has been treated separately, this method is range-independent. (Note that

IMPORTANCE[] is an array of predefined "importance" weights.)

Algorithm 3 Weighted Average Ranking (WAR)

for every objective in problem **do**
  Form a list of the fitness of each solution and pointer to this solution
  for current objective **do**
    Sort $FITNESS - LIST$ into order of fitness
  end for
end for
Set every $SOLUTION.RANK = 0$
for every ranking position P in population **do**
  for every objective O in problem **do**
    ** $FITNESS-LIST$ for $O[P]- > SOLUTION.RANK+ = P*IMPORTANCE$
  end for
end for

Corne and Knowles [3] 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. It should be noted that in their AR method, the value of importance array of "IMPORTANCE[]", mentioned above, is set to one.

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.

Chapter 6

Decoding

The *decoding* method proposed by Schliep et al. [43], 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 $\{p_1,...,p_n\}$ as the solution of non-unique probe selection, and a result vector $r = (r_1,...,r_n)$ 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:

$$P[S|r] = \frac{P[r|S].P[S]}{P[r]} \tag{6.1}$$

$P[r|S]$ 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. [43] have defined the $P[r|S]$ as:

$$P[r|S] = \prod_{p_j} f(r_j, |S(j) \cap S|), \tag{6.2}$$

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

$$P[S] \propto c_{|S|}. \prod_{t_i \in S} f_i \prod_{t_i \notin S} (1 - f_i) \tag{6.3}$$

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 subsets of T that include targets t.

$$P[t_i|r] \propto \sum {}_{S:t \in S} P[S|r] \tag{6.4}$$

Since $P[r]$ is not available, the posterior can not 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. [43] 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 : $pos_1, pos_2, ..., pos_l$. The objective Obj_{dec} is defined as following:

$$Obj_{dec} = \frac{1}{\sum_{i=1}^{l} pos_i} \tag{6.5}$$

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: $\frac{(l-1)\times(l)}{2}$. We examine at most 100 targets in the top 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 first 100 positions of the targets ranked list.

Chapter 7

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, 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 [15] 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 [54].

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 3.2. 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. In all the experiments, we set the variable k to two. According to the experiments which will be explained in section 7.2.2, increasing the dependency parameter did not result in better probe sets [45]. This parameter is equal to the maximum number of incoming edges to each node of the Bayesian Netwrok used in the BOA software [55] 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.

7.1 Data Sets

We have performed the experiments on ten artificial datasets called a1,..., a5, b1,..., b5, and two real datasets HIV1 and HIV2. All previous studies mentioned in section 2 have been conducted on the same datasets, except for the HIV1, and HIV2 that have not been used in [22] [21]. As mentioned, the datasets are the target-probe incidence matrices. Properties of the datasets are presented in Table 7.1. Along with this information, the number of virtual probes required for each dataset has been noted. 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.

The artificial datasets a1,...,a5,b1,...,b5 has been generated by means of Random Evolutionary FORest Model (REFORM) software [39]. Ten first test sets of 256 targets (a1,...,a5) have been generated by one model, and the next five sets (b1,...,b5) with 400 targets have been generated by another model. for further information on the sets generation, see [22].

The sets of HIV1 and HIV2 with 200 targets sequences for each have been downloaded from the National Center for Biotechnology Information (NCBI). These datasets contain similar sequences that make them appropriate sets for the non-unique probe selection problem. The candidate probes of these sets have been generation by means of Primer3 software. The input parametes used for this software are: probe length between 18 and 27 nucleotides, melting temperature between 57,° C and 63,° C, and GC content between 20 and 80%. By the software, 40 probes for each HIV sequence (eight thousand in total), were generated

Table 7.1: Properties of the datasets used for experiments. 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.

Set	$\|T\|$	$\|P\|$	$\|V\|$
a1	256	2786	6
a2	256	2821	2
a3	256	2871	16
a4	256	2954	2
a5	256	2968	4
b1	400	6292	0
b2	400	6283	1
b3	400	6311	5
b4	400	6223	0
b5	400	6285	3
HIV1	200	4806	20
HIV2	200	4686	35

for each dataset. Before constructing the HIV target-probe incidence matrices, the repeat probes have been filtered [35].

7.2 Single targets in sample

7.2.1 Experiments with the default parameters:

First series of experiments have been performed with the default parameters of BOA [55]. 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 7.2. 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 re-

Table 7.2: 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 Equation 7.1).

Set	$\|T\|$	$\|P\|$	$\|V\|$	ILP[22][21]	OCP[35]	DRC[51] -GA	BOA +SDRC	BOA +DPS	BOA +DRC
a1	256	2786	6	503	509	502	503	503	502
a2	256	2821	2	519	494	490	492	491	490
a3	256	2871	16	516	543	534	535	533	533
a4	256	2954	2	540	539	537	540	538	537
a5	256	2968	4	504	529	528	530	530	528
b1	400	6292	0	879	830	839	843	837	834
b2	400	6283	1	938	842	852	853	849	846
b3	400	6311	5	891	827	835	839	831	829
b4	400	6223	0	915	873	879	877	877	875
b5	400	6285	3	946	874	890	887	886	879
HIV1	200	4806	20	-	451	450	452	450	450
HIV2	200	4686	35	-	479	476	479	475	474

spectively 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) [22][21], Optimal Cutting Plane Algorithm (OCP) [35], and Genetic Algorithm (DRC-GA) [51]. The last three columns show the improvement of our approach over each of the three latest approaches. The improvement is calculated by Equation 7.1.

$$Imp = \frac{P_{min}^{BOA+DRC} - P_{min}^{Method}}{P_{min}^{Method}} \times 100 \tag{7.1}$$

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

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 7.3, 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.

Table 7.3: The last three columns are showing the improvement of BOA+DRC over three methods ILP, OCP, and DRC-GA (see Equation 7.1)

Set	ILP[22][21]	OCP[35]	DRC-GA[51]
a1	-0.20	-1.37	0
a2	-5.59	-0.81	0
a3	+1.35	-2.02	-0.18
a4	-0.55	-0.37	0
a5	+4.76	-0.19	0
b1	-5.12	+0.50	-0.60
b2	-9.81	+0.47	-0.70
b3	-6.96	+0.24	-0.72
b4	-4.37	+0.23	-0.45
b5	-7.08	+0.57	-1.23
HIV1	-	-0.22	0
HIV2	-	-1.04	-0.42

As shown in the Table 7.2, 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 [52]. The results obtained in the [35] are considered as the best ones in the literature for the non-unique probe selection problem. As shown in the 7.2, Wang et. al. [51] have recently reported the results (noted as DRC-GA) which are comparable to (and in most cases better than) [35].

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 7.4.

Table 7.4: 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 7.2) 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

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 to, also have correlations with each other.

A part of the obtained dependencies between probes for dataset HIV2 is presented in Figure 7.1. Network display of this output is demonstrated in Figure 7.2. 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.

7.2.2 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

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

Figure 7.1: Part of the BOA output for dataset HIV2: the discovered dependencies for probes 30 to 38 by BOA.

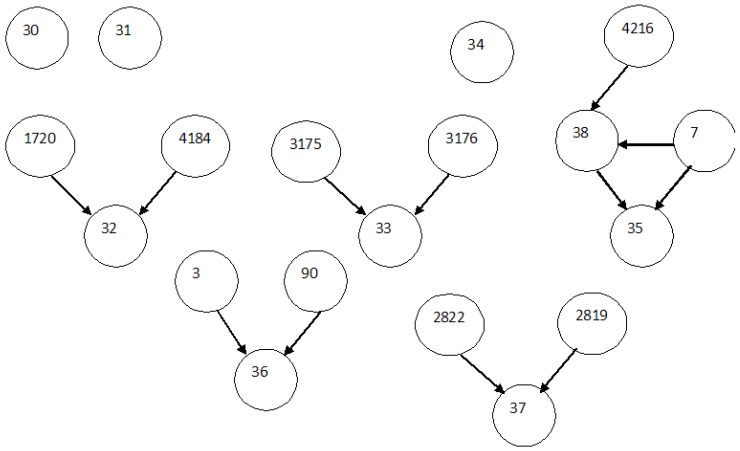

Figure 7.2: Network demonstration of the BOA output from Figure 7.1

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 7.6.

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 7.5: 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

7.3 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 [46]. We applied the multiobjective optimization technique presented in section 5, and measured the ability of the probe set in 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 6.

The experiments were conducted in two main series of identification of five and ten targets, and identification of fifteen and twenty 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.

7.3.1 Identification of five and ten targets

In the first series of experiments, the goal was set to measure the ability of the solutions in simultaneously identifying five and ten targets in the sample. The results are presented in the table 7.6.

As mentioned, first, 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

Table 7.6: 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)
a1	508	515
a2	494	502
a3	537	545
a4	540	546
a5	533	539
b1	867	879
b2	883	897
b3	864	872
b4	891	912
b5	920	938
HIV1	456	458
HIV2	483	487

selected probe set (in first iteration of BOA) is able to identify five targets in the sample for all the datasets.

As presented in the Table 7.6, 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 7.2). 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 is this case, our observation was similar to the

previous step. As mentioned before, we have set the separation constraint (s_{min}) to five. By applying the DRC heuristic (4.2) 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 problems of optimization difficult enough. 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 7.3.2).

7.3.2 Identification of fifteen and twenty targets

Since the obtained probe sets by our method had a high ability to identify multiple (five and ten) targets in the sample, we tried to increase the number of targets in the sample, and 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 fifteen and twenty targets in the sample.

We conducted new experiments for all the datasets. Table 7.7 indicates the cardinality of the minimal probe sets obtained for datasets a1,...,a5 in the new experiments. As mentioned before, the obtained probe sets by multiobjective optimization are larger than the obtained probe sets by one-objective optimization problem.

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

Table 7.7: Cardinality of minimum probe set obtained by applying the BOA+DRC in case of multiple targets in the sample - two cases of fifteen and twenty targets in the sample were considered.

Set	BOA+DRC (15 targets)	BOA+DRC (20 targets)
a1	517	524
a2	504	507
a3	549	553
a4	548	552
a5	544	547

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 fifteen targets) in 40 iterations of BOA for dataset $a3$ is shown in Figure 7.3.

In Figure 7.3, the maximum decoding score obtained in each iteration of BOA is presented. The maximum possible decoding score for the case of fifteen targets is obtained when all the targets are identified by the probe set as the top fifteen positions. Therefore, the value of the maximum score is $\frac{1}{105} \approx 0.009524$. As shown in 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 6, the inverse of the maximum decoding score in case of fifteen 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 (Equation 7.2).

$$AverageTargetsPosition = \frac{\sum_{i=1}^{l} pos_{t_i}}{l} \quad 1 \leq i \leq l \tag{7.2}$$

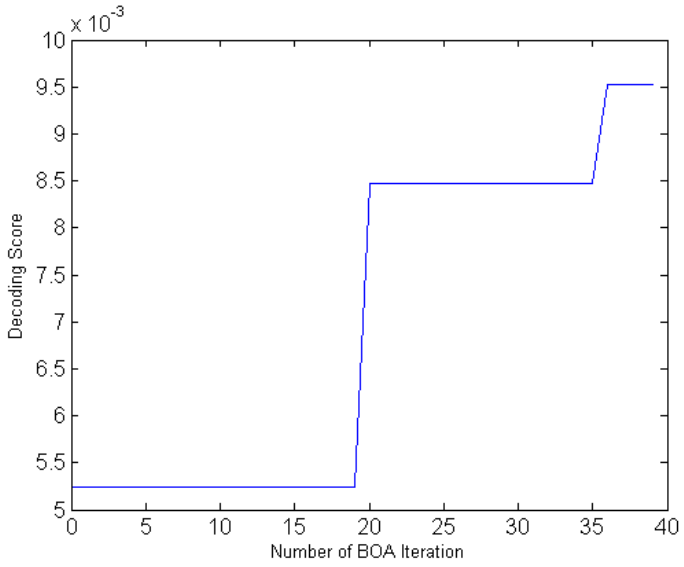

Figure 7.3: Maximum decoding score for dataset $a3$ in 40 iterations of multiobjective optimization in case of fifteen targets in the sample.

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 7.2), and averaged the score over 50 runs for each of the datasets. 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 is demonstrated in the Table 7.8. For instance, the average target position identified by the optimal probe set obtained in case of single target in sample, for dataset $a3$, is 49.93. By applying multiobjective optimization method, we have improved this value to its best possible value (7) in case of fifteen 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, during

40 iterations, the decoding score has not been improved considerably for the datasets a1, a2, and a5. The problem of identifying fifteen targets in the sample can be considered a difficult problem for these datasets, and further attempts are required in order to solve these problems more efficiently.

The same calculations can be conducted for the case of twenty targets in the sample (see Table 7.9). The maximum decoding score in this case is $\frac{1}{190} \approx 0.005263$. 190 which is the summation of twenty targets positions results in $\frac{190}{15} \approx 12.67$ average target position for this case.

As presented in the Table 7.9, comparing with the optimal probe set obtained by the one-

Table 7.8: 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 multiobjetcive optimization in case of fifteen targets in the sample. The average target position (Ave Target Position) corresponding to each score is also presented. Maximum possible decoding score (0.009523) has been obtained for datasets $a3$, $a4$, $b1$, $b2$, $b3$, $b4$, $b5$, and almost for $HIV2$.

Set	Ave Dec Score	Ave Target position	Max Dec Score	Ave Target Position
a1	0.001300	51.28	0.005235	12.73
a2	0.001304	51.12	0.005235	12.73
a3	0.001335	49.93	**0.009523**	**7**
a4	0.001338	49.82	**0.009523**	**7**
a5	0.001218	54.73	0.005235	12.73
b1	0.001499	44.47	**0.009523**	**7**
b2	0.001486	44.86	**0.009523**	**7**
b3	0.001477	45.14	**0.009523**	**7**
b4	0.001627	40.97	**0.009523**	**7**
b5	0.001476	45.17	**0.009523**	**7**
HIV1	0.000956	69.73	0.003597	18.53
HIV2	0.001196	55.74	**0.009346**	**7.13**

Table 7.9: 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 multiobjetcive optimization in case of twenty targets in the sample. 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 dadaset $b3$ and almost $b4$.

Set	Ave Dec Score	Ave Target Position	Max Dec Score	Ave Target Position
a1	0.000920	54.35	0.002747	18.20
a2	0.000898	55.68	0.002695	18.55
a3	0.000885	56.50	0.002824	17.70
a4	0.000988	50.61	0.002808	17.81
a5	0.000828	60.39	0.002293	21.80
b1	0.000989	50.56	0.002391	20.91
b2	0.001067	46.86	0.003690	13.55
b3	0.001177	42.48	**0.005236**	**9.54**
b4	0.001152	43.40	**0.005263**	**9.5**
b5	0.001037	48.22	0.003690	13.55
HIV1	0.000677	73.85	0.002062	24.24
HIV2	0.001134	44.09	0.002732	18

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 twenty 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 of 'maximum incoming edges' on the decoding objective. The maximum incoming edges, (see section 3.2), determines the level of dependency among variables in BOA.

Comparison between optimized and random solutions of same length

Following the experiments illustrated in section 7.3.2, we performed another series of interesting experiments on the dataset $a3$, 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 multiobjective 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 the Table 7.10, the minimum length obtained in case of single target in the sample (experiments of section 7.2, Table 7.2) for some datasets are demonstrated. Along with these, the length of the solution with the maximum decoding value in case of twenty targets in the sample is indicated for mentioned datasets.

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 building a set of the same cardinality mentioned in the third column of the Table 7.10. Then, the decoding score of the resulted probe set, for each dataset, was compared with the obtained maximum decoding score in the case of twenty targets. The result is illustrated in Table 7.11.

As noted in Table 7.11, 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 twenty targets in the sample is shown. Considerable increase obtained, by applying optimization algorithm, can be noticed by comparing these two values for each dataset.

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,

Table 7.10: 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 twenty targets in the sample.

Set	Minimum Length (single target in sample)	Length (of the solution with maximum decoding score)
a3	533	681
b1	834	968
b2	846	989
b3	829	932
b4	875	1159
b5	879	1010
HIV1	450	525
HIV2	474	584

Table 7.11: Comparing the decoding ability of the optimized solution in case of twenty targets in the sample to the decoding ability of a random solution of the same length.

Set	Random Solution	Optimized Solution
a3	0.000869	0.002824
b1	0.000893	0.002391
b2	0.000909	0.003690
b3	0.001047	0.005236
b4	0.001094	0.005263
b5	0.001010	0.003690
HIV1	0.000674	0.002062
HIV2	0.000778	0.002732

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.

Chapter 8

Conclusions

8.1 Summary of Contributions

In this thesis, we presented a new approach for solving the non-unique probe selection problem. Our approach is based on the combination of one of the EDAs named BOA with the simple and fast heuristics proposed for solving the non-unique probe selection problem. We obtained 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. In the case of single target in the sample, 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.

Moreover, for the case of multiple targets in the sample, we applied an extended version of the combination of BOA and DRC. We considered a second objective for the problem which was the ability of the selected probe set in identification of multiple targets in the sample. By applying a modified version of the decoding (chapter 6), we tried to measure the ability of the solutions in achieving the second goal. Our work is the first one that

explicitly considers the decoding ability as an objective for the optimization problem.

Our goal was to approach the non-unique probe selection problem in case of multiple targets as a two-objective optimization problem. We conducted the experiments in case of five and ten targets in the sample. 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. Therefore, 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 fifteen and twenty targets in the sample constructed difficult optimization problems. Our method was successful in solving the optimization problem for the case of fifteen targets for the datasets a3 and a4, and 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 fifteen and twenty targets in the sample. Moreover, we believe that our multiobjective-based method makes a flexible framework for the problem of non-unique probe selection.

8.2 Future Work

As mentioned in the experiments section 7.2.2 related to the one-objective problem or the case of single target in the sample, we investigated the effect of increasing the dependencies among variables discovered by BOA for some of the datasets. According to the presented results, it is possible that the minimal probe sets found for some of these datasets are the *global* optimal values. This is a subject that requires more experiments and investigation

in future.

Also, one specific advantage of our approach is discovering the dependencies among the variables or probes. These discovered dependencies can be interpreted more precisely by biologists in order to detect more interesting information about the relation between probes and the targets to which they hybridize.

In the case of multiple targets in the sample, we are interested in examining the impact of modification in the BOA parameters on the decoding ability of the solutions. For instance, the impact of increasing the maximum number of dependencies between the variables on the decoding ability can be investigated in further studies.

Moreover, as mentioned in the experiments section related to multiple targets, the experiments are performed for 40 iterations of BOA. The possibility of improvement in decoding ability of the solutions by increasing the number of iterations should be studied.

On the other hand, we believe that our extended approach for the case of multiple targets is very flexible. Hence, in further studies, it will be interesting to consider new objectives and integrate them to the optimization 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.

References

[1] P. J. Bentley and J. P. Wakefield, "Finding Acceptable Solutions in the Pareto-Optimal Range using Multiobjective Genetic Algorithms", Second Online World Conference on Soft Computing in Engineering Design and Manufacturing (WSC2) 5, p. 242-249, 1998.

[2] J. Borneman, M. Chroback, G.D. Vedona, A. Figueroa, and T. Jiang. "Probe Selection Algorithms with Applications in the Analysis of Microbial Comunities". *Bioinformatics 17*, Suppl 1: s39-s48, 2001.

[3] D. W. Corne, J. D. Knowles, "Techniques for Highly Multiobjective Optimization: Some Non-dominated Points Are Better than Others", GECCO 2007: 773-780.

[4] A. Cutichia, J. Arnold, and W. Timberlake. "PCAP: Probe Choice and Analysis Package - Aset of Programs to Aid in Choosing Systhetic Oligomers for Contig Mapping". *CABIOS 9*, 201-203, 1993.

[5] D. J. Cutler, M. E. Zwick, M. M. Carrasquillo, C. T. Yohn, K. P. Tobin, C. Kashuk, D. J. Mathews, N. A. Shah, E. E. Eichler, J. A. Warrington, and A. Chakravarti. "High-throughput variation detection and genotyping using microarrays". Genome Research, 11(11):19131925, 2001.

[6] P. Deng, F. Wang, and D.Z. Du. "Non-unique Probe Selection with Group Testing". *Proceeding of the First International Symposium on Optimization and System Biology (OSB'07)*, Beijing, China, 2007.

[7] P. Deng, M.T. Thai, Q. Ma, and W. Wu. "Efficient Non-unique Probes Selection Algorithms for DNA Microarray." *BMC Genomics 9*, Suppl 1:S22, 2008.

[8] D. Z. Du and F. K. Hwang. Combinatorial Group Testing and Applications. World Scientific Publishing, 1993.

[9] Y. X. Fu, W. E. Timberlake, and J. Arnold. "On the Design of Genome Mapping Experiments Using Short Synthetic Oligonucleotides", *Biometrics*, 48, 337-359, 1992.

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

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

[12] L. Gasieniec, C.Y. Li, P. Sant, and P.W.H. Wong. "Efficient Probe Selection in Microarray Design." *Proceeding of IEEE Symposium on Computational Intelligence and Bioinformatics and Computationalbiology (CIBCB'06)*, Toronto, Ontario, Canada, 2006.

[13] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, et al. "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.", Science, 286(5439):531537, 1999.

[14] R. Gras, "How Efficient Are Genetic Algorithms to Solve High Epistasis Deceptive Problems?", 2008 IEEE Congress on Evolutionary Computation, June 1-6, Hong Kong, China, p. 242-249, 2008.

[15] W. Gropp, E. Lusk, and A. Skjellum, Using MPI: Portable Parallel Programming with the Message-passing Interface. MIT Press In Scientific And Engineering Computation Series, MA, USA, 1994.

[16] J. G. Hacia, L. C. Brody, M. S. Chee, S. Fodor, and F. S. Collins. "Detection of heterozygous mutations in BRCA1 using high density oligonucleotide arrays and two-colour fluorescence analysis.", Nature Genetics, 14(4):441447, 1996.

[17] D. Heckerman, D. Geiger, and D. Chickering, "Learning Bayesian Networks: The Combination of Knowledge and Statistical Data", Tenth Conference on Uncertainty in Artificial Intelligence, Seattle, WA, p. 293301, 1994.

[18] R. Herwig, A.O. Schmitt, M. Steinfath, J. O'Brien, H. Seidel, S. Meier-Ewert, H. Lehrach, and H. Randelof. "Information Theoretical Probe Selection for Hybridization Experiments.", *Bioinformatics 16*, 10, 890-898, 2000.

[19] D. S. Hochbaum, Approximation Algorithms for NP-hard Problems., PWS publishing, 1997.

[20] T. R. Hughes, M. Mao, A. R. Jones, J. Burchard, M. J. Marton, K. W. Shannon, S. M. Lefkowitz, M. Ziman, J. M. Schelter, M. R. Meyer, S. Kobayashi, C. Davis, H. Dai, Y. D. He, S. B. Stephaniants, G. Cavet, W. L. Walker, A. West, E. Coffey, D. D. Shoemaker, R. Stoughton, A. P. Blanchard, S. H. Friend, and P. S. Linsley. "Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer.", Nature Biotechnology, 19(4):342347, 2001.

[21] G. W. Klau, S. Rahmann, A. Schliep, M. Vingron, and K. Reinert, "Integer Linear Programming Approaches for Non-unique Probe Selection", *Discrete Applied Mathematics*, vol. 155, pp. 840–856, 2007.

[22] G. W. Klau, S. Rahmann, A. Schliep, M. Vingron, and K. Reinert, "Optimal Robust Non-unique Probe Selection Using Integer Linear Programming", *Bioinformatics*, vol. 20, pp. i186–i193, 2004.

[23] M. Kozal, N. Shah, N. Shen, R. Yang, R. Fucini, T. C. Merigan, D. D. Richman, D. Morris, E. Hubbell, M. Chee, and T. R. Gingeras. "Extensive polymorphisms observed in HIV-1 clade B protease gene using high-density oligonucleotide arrays." Nature Medicine, 2(7):753759, 1996.

[24] P. Larrañaga and J. A. Lozano, Estimation of Distribution Algorithms: A New Tool for Evolutionary Computation. Kluwer Academic Publishers, 2001.

[25] G. G. Lennon and H. Lehrach. "Hybridization analyses of arrayed cDNA libraries.", Trends in Genetics, 7(10):314317, 1991.

[26] R. J. Lipshutz et al. "Advanced DNA sequencing technologies.", Current Opinion in Structural Biology, 4:376380, 1994.

[27] F. Li, and G. Stormo. "Selection Optimum DNA Oligos for Microarrays.", *Proceedings of IEEE International Symposium on Bio-Informatics and Biomedical Engineering (BIBE)*, Key Bridge Marriot, Arlington, USA, 2000.

[28] D.J. Lockhart, H. Dong, M.C. Byrne, M.T. Follettie, M.V. Gallo, M.S. Chee, M. Mittmann, C. Wang, M. Kobayashi, H. Horton, and E.L. Brown. "Expression Monitoring by Hybridization to High-Density Oligonucleotide Arrays." *Nature Biotechnology 14*, 13, 1675-1680, 1996.

[29] F. Lottspeich and H. Zorbas, editors. Bioanalytik. Spektrum Akademischer Verlag, 1998.

[30] C. N. Meneses, P. M. Pardalos, and M. A. Ragle, "A New Approach to the Non-unique Probe Selection Problem", *Annals of Biomedical Engineering*, vol. 35, no. 4, pp. 651–658, 2007.

[31] O. Milenkovic, and N Kashyap. "DNA Codes that Avoid Secondary Structure.", *Proceeding of the 2005 IEEE International Symposium on Information Theory.*, Adelaide, Australia, 2005.

[32] R. Murray, D. Granner, P. Mayes, and V. Rodwell. Harper's Illustrated Biochemistry (26th ed). McGraw-Hill Companies, 2004.

[33] H. Mühlenbein and G. Paaß, "From Recombination of Genes to the Estimation of Distributions I. Binary Parameters", 4th International Conference on Parallel Problem Solving from Nature, p. 178–187, September 22-26, 1996.

[34] M. Pelikan, Bayesian Optimization Algorithm: From Single Level to Hierarchy. University of Illinois. PhD Thesis, 2002.

[35] M. A. Ragle, J. C. Smith, and P. M. Pardalos, "An Optimal Cutting-plane Algorithm for Solving the Non-unique Probe Selection Problem", *Annals of Biomedical Engineering*, vol. 35, no. 11, pp. 2023–2030, 2007.

[36] S. Rahman. "Rapid Large-Scale Oligonucleotide Selection fpr Microarrays.', *Proceeding of the First IEEE Computer Society Bioinformatics Conference (CSB)*, 54-63, Stanford, CA, USA, 2002.

[37] S. Rahman, T. Muller, and M. Vingron. "Non-unique Probe Selection by Matrix Condition Optimization". *Currents in Computational Molecular Biology*, San Diego, USA, 2004.

[38] S. Rahman. Algorithms for Probe Selection and DNA Microarray Design. Dissertation. Max Plank Institute for Molecular Genetics, Berlin, 2004.

[39] S. Rahmann. REFORM (Random Evolutionary FORests Modeling software), 2003.

[40] S. Rash, D. Gusfield, "String Barcoding: Uncovering Optimal Virus Signatures", *Annual Conference on Research in Computational Molecular Biology*, p. 254–261 , 2002.

[41] S. Rozen and H. Akaletsky Primer3 on the WWW for general users and for biologyst programmers. *Bioinformatics Methods and Protocols: Methods in Molecular Biology*, edited by S. Krawetz and S. Misener. Totowa, NJ: Humana Press, 365-386, 2000.

[42] J. SantaLucia, H. Allawi, and P. Seneviratne. "Improved Nearest-Neighbor Parameters for Predicting DNA Duplex Stability.", *Biochemistry 35*:3555-3562, 1996.

[43] A. Schliep, D. C. Torney, and S. Rahmann, "Group Testing with DNA Chips: Generating Designs and Decoding Experiments", IEEE Computer Society Bioinformatics Conference (CSB'03), p. 84–91, 2003.

[44] J. D. Schaffer, "Multiple Objective Optimization with Vector Evaluated Genetic Algorithms", Gentic Algorithms and Their Applications: 1st International Conference on Genetic Algorithms, p. 93–100, 1985.

[45] L. Soltan Ghoraie, R. Gras, L. Wang, A. Ngom, "Bayesian Optimization Algorithm for the Non-unique Oligonucleotide Probe Selection Problem.", In Proceedings of the fourth IAPR International Conference on Pattern Recognition in Bioinformatics, Sheffield, UK, 365-376, 2009.

[46] L. Soltan Ghoraie, R. Gras, L. Wang, A. Ngom, "Optimal Decoding and Minimal Length for the Non-unique Oligonucleotide Probe Selection Problem.", 2009. (Submitted to the journal of Neurocomputing, special issue on Bioinformatics).

[47] D. Stekel. Microarray Bioinformatics. Cambridge University Press, Cambridge.

[48] T. Strachan and A. P. Read. Human Molecular Genetics. Garland Science Publishers, 3rd edition, 2003.

[49] W.K. Sung, and W.H. Lee. "Fast and Accurate Probe Selection Algorithm for Large Genomes", *Proceeding of the 2nd IEEE Computer Society Bioinformatics Conference (CSB'03)*, Stanford, CA, USA, 2003.

[50] M. Thai, D. MacCallum, P. Deng, and W. Wu. "Decoding Algorithms in Pooling Designs with Inhibitors and Error-Tolerance.", *Int. J. Bioinformatics Research and Applications 3*, 2, 145-152, 2007.

[51] L. Wang, A. Ngom, and R. Gras, "Non-Unique Oligonucleotide Microarray Probe Selection Method Based on Genetic Algorithms", 2008 IEEE Congress on Evolutionary Computation, June 1-6, Hong Kong, China, p. 1004–1010, 2008.

[52] L. Wang, and A. Ngom, "A Model-based Approach to the Non-unique Oligonucleotide Probe Selection Problem", Second International Conference on Bio-Inspired Models of Network, Information, and Computing Systems (Bionetics 2007), December 10-13, Budapest, Hungary, ISBN: 978-963-9799-05-9, 2007.

[53] L. Wang, A. Ngom, R. Gras and L. Rueda, "Evolution Strategy with Greedy Probe Selection Heuristics for the Non-unique Oligonucleotide Probe Selection Problem", 2008 IEEE Symposiunm on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB 2008), p. 54–61, 2008.

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

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

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

[57] http://www.repeatmasker.org/

[58] ILOG, Inc. CPLEX. http://www.ilog.com/products/cplex, 19872004.

[59] http://www.cs.umsl.edu/ pelikan/boa.html

VITA AUCTORIS

Laleh Soltan Ghoraie was born in 1983 in Tehran, Iran. In 2005, he received his Bachelor of Applied Science degree in Computer Engineering from University of Shahid Beheshti (National University of Iran) in Tehran, Iran. In 2007, she went to Canada to pursue graduate studies. She has been working as a research assistant in the Pattern Recognition and Machine Learning Lab under the supervision of Doctor Robin Gras (CRC in Probabilistic Model Building in Bioinformatics) for the last two years.