Multi-objective Evolutionary Probe Design Based on Thermodynamic Criteria for HPV Detection

In-Hee Lee, Sun Kim, and Byoung-Tak Zhang

Biointelligence Laboratory
School of Computer Science and Engineering
Seoul National University, Seoul 151-742, Korea
{ihlee,skim,btzhang}@bi.snu.ac.kr

Abstract. DNA microarrays are widely used techniques in molecular biology and DNA computing area. It consists of the DNA sequences called probes, which are DNA complementaries to the genes of interest, on solid surfaces. And its reliability seriously depends on the quality of the probe sequences. Therefore, one must carefully choose the probe sets in target sequences. In this paper, the probe design for DNA microarrays is formulated as the multi-objective optimization problem. We propose a multi-objective evolutionary approach, which is known to be suitable for this kind of optimization problem. Since a multi-objective evolutionary algorithm can find multiple solutions at a time, we used thermodynamic criteria to choose the most suitable one. For the experiments, the probe set generated by the proposed method is compared to the sequences used in commercial microarrays, which detects a set of Human Papillomavirus (HPV). The comparison result supports that our approach can be useful to optimize probe sequences.

Contents Area: Bioinformatics and AI, Evolutionary computing

1 Introduction

DNA microarray is a small plate on which various kinds of oligonucleotide probes are attached. It is widely used to study cell cycle, gene expression profiling and other DNA-related phenomena in a cell. When the contents of a cell is hybridized to the microarray, if there exists a complementary molecule to one of the probes, it would hybridize to the probe so that a user can detect it. In this way, it can provide the information on whether a gene is expressed or not for hundreds of genes at a time. From this, a biologist can get an overview of gene expression level at a certain time point.

There are two kinds of DNA microarray, cDNA microarray and oligonucleotide microarray. In contrast to cDNA microarrays, one can choose or change the probe sequences on oligonucleotide microarrays. Therefore, the reliability of the information that a oligonucleotide microarray provides depends on the quality of probe sets that used. If a probe hybridizes to not only its target gene but

C. Zhang, H.W. Guesgen, W.K. Yeap (Eds.): PRICAI 2004, LNAI 3157, pp. 742–750, 2004.

also other genes, the microarray may produce misleading data. Thus, one needs to design the probe set carefully to get precise data.

In literature, lots of probe design methods are suggested reflecting its importance. In [4], the frequency matrix based method was suggested. And an information theoretical method based on Shannon entropy as a quality criterion was used in [5]. Li *et al.* suggested a method based on sequence information and hybridization free energy in [7]. The optimum probes are picked based on having free energy for the correct target, and maximizing the difference in free energy to other mismatched target sequences. And Bourneman *et al.* proposed two heuristics for minimizing the number of oligonucleotide probes needed for analyzing populations of ribosomal RNA gene clones on DNA microarrays [1]. One was a simulated annealing based method which was used to find the probe sets maximizing the number of distinguished pairs of clones for the given number of probes, and the other heuristic, the Lagrangian relexation, was applied to find a minimum number of probes that distinguish all given clones. Recently, a method based on machine learning algorithms such as naïve Bayes, decision trees, and neural networks has been also proposed for aiding probe selection [9]. It tests the probe sets which has high possibility for the hybridization experiments by the learning on the microarray data from *E. coli* and *B. subtilis*. But in spite of the variety of probe design methods, there exists no evolutionary computation-based approach.

We formulated the problem of selecting optimal set of probes as multi-objective optimization problem and applied multi-objective evolutionary algorithm. A multi-objective optimization problem usually has a set of Pareto optimal solutions instead of only one optimal solution. The multi-objective evolutionary algorithm has the advantage that one can get the Pareto optimal solutions at a time. But in the real-world application, one must choose one solution rather than the set of Pareto optimal solutions. As a final decision maker, we choose the thermodynamic criteria for it can provide a realistic evaluation of the set of probes.

In the following sections, we explain the suggested probe design method in detail. In section 2, we briefly introduce the multi-objective optimization problem and formulates the probe design problem as multi-objective optimization problem. Section 3 and 4 describe our probe design method and provide the experimental results. Conclusions are drawn in section 5.

2 Multi-objective Formulation of Probe Design

2.1 Multi-objective Optimization Problem

As the name suggests, a multi-objective optimization problem (MOP) has a number of objectives which are to be optimized [3]. And the problem usually has a number of constraints. The general form of multi-objective optimization problem is like the following:

$$\begin{aligned} \text{Optimize} \quad & f_i(\mathbf{X}), \quad i = 1, \cdots, M; \\ \text{subject to} \quad & g_j(\mathbf{X}) = 0, \, j = 1, \cdots, N. \end{aligned}$$

Here, M denotes the number of objectives and N the number of constraints. And a solution can be represented as a vector in the objective space, denoted by $f(\mathbf{X}) = (f_1(\mathbf{X}), f_2(\mathbf{X}), \ldots, f_M(\mathbf{X}))$.

We suppose conflicting objectives and no priority between objectives in the further explanation. For non-conflicting objectives, the optimization of one objective implies the optimization of the other and both objectives can be treated as one objective. And if there exists priority between objectives, one can optimize objectives according to the priority by optimizing single objective which is the weighted sum of objectives. Therefore, for both cases, the given problem becomes a single objective optimization problem and we exclude such cases hereafter.

Given an optimization problem, one's goal is to find optimal solution(s). For a single objective case, the optimality of a solution is determined by simply comparing its objective function value to others. In multi-objective case, the optimality of a solution is determined by domination relation between solutions. A solution X is said to *dominate* other solution Y when the following two conditions are satisfied and denoted by $X \preceq Y$:

1. X is no worse than Y in all objectives.
2. X is strictly better than Y in at least one objective.

Therefore, the optimal solutions for a MOP are those that are not dominated by any other solutions. Thus, one's goal in MOP is to find such a *non-dominated set* of solutions.

There exist several methods to find such non-dominated set of solutions for an MOP. Among them, evolutionary method is most popular and currently most actively studied method. Because it is a population-based method, it has the advantage that it can provide a set of non-dominated solutions by one run.

2.2 Probe Design as an MOP

There exist several criteria to evaluate the set of probes. We list the generally used conditions for good probes as follows:

1. The probe sequence for each gene should not appear other genes except its target gene.
2. The non-specific interaction between probe and target should be minimized.
3. The probe sequence for each gene should be different from each other as much as possible.
4. The probe sequence for each gene should not have secondary structure such as hairpin.
5. The melting temperatures of the probes should be uniform.

The first three conditions concern with the specificity of the probes. And the secondary structure of a probe can disturb the hybridization with its target gene. Therefore, well-designed probes should have minimal secondary structure. Lastly, the probes on a oligonucleotide chip is exposed to the same experimental

condition. If the melting temperatures of the probes are not uniform, some probes can not hybridize with its target. So, the probes must have the uniform melting temperatures.

Before going on the formulation of the problem, let us introduce some notations. We denote a set of n probes by $X = \{x_1, x_2, \ldots, x_n\}$, where $x_i = \{A, C, G, T\}^l$ for $i = 1, 2, \ldots, n$, l is the length of each probe. And we denote the set of target genes by $T = \{t_1, t_2, \ldots, t_n\}$.

The first condition is the basic requirement for a valid set of probes. Therefore it can be treated as a constraint as follows:

$$g(X) = \sum_{i \neq j} subsequence(x_i, t_j),$$

where $subsequence(x_i, t_j)$ is one if x_i occurs in t_j at least once and zero otherwise. From its definition, this constraint must be zero.

Other 4 conditions can be formulated as minimization objectives. These are formulated as follows:

$$\begin{aligned} f_1(X) &= \sum_{i \neq j} hybridize(x_i, t_j), \\ f_2(X) &= \sum_{i \neq j} similarity(x_i, x_j), \\ f_3(X) &= \sum_{i} hairpin(x_i), \\ f_4(X) &= \sigma_{Tm(X)}. \end{aligned}$$

Here, $hybridize(x_i, t_j)$ has non-zero value in proportion to the hybridization likelihood between x_i and t_j. And $similarity(x_i, x_j)$ means the hamming distance between x_i, x_j including shifted comparison. For condition 4, we considered hairpin as the only possible secondary structure, because other structures are hard to compute or hardly occur in microarray probes. $hairpin(x_i)$ has non-zero value in accordance with the probability that x_i can form a hairpin. Condition 5 can be formulated as minimizing the standard deviation of melting temperatures of each probes ($Tm(X)$).

From above, the probe design problem is formulated as an MOP with 4 minimization objectives and 1 equality constraints.

$$\begin{aligned} \text{Minimize} \quad & f_i(\mathbf{X}), \quad i = 1, \cdots, 4; \\ \text{subject to} \quad & g(\mathbf{X}) = 0\,. \end{aligned}$$

3 Evolutionary Oligonucleotide Probe Design

To design probe set that satisfies above condition, we propose an evolutionary computation-based approach. In this approach, we try to find various optimal solutions simultaneously using multi-objective evolutionary algorithm and then choose appropriate probe set according to thermodynamic criteria.

In the following subsections, we describe the algorithm and the thermodynamic criteria in detail.

3.1 The Multi-objective Evolutionary Approach

We try to find optimal probe sets using multi-objective evolutionary algorithm. For the multi-objective evolutionary algorithm produces multiple solutions, we choose the most appropriate solution using thermodynamic criteria.

When we optimize probe sets, we applied multi-objective evolutionary algorithm for the following reasons. First, the probe design can be viewed as a kind of multi-objective optimization problem, and evolutionary algorithm showed good performance in optimization including multi-objective optimization. Second, evolutionary approach can provide a population of solutions rather than one solution. The probe design is a multi-objective optimization between conflicting objectives. For example, if we extremely maximize the difference between the probe sequences, the specificity of the probes may drop. Therefore, each set of probes is a trade-off between these conflicting objectives. Evolutionary multi-objective algorithm can find multiple trade-off solutions of various degree. Users can find the most suitable trade-off solution among the population according to appropriate criteria. In this paper, we choose the thermodynamic criteria.

We used the multi-objective evolutionary algorithm called NSGA-II [2]. It is based on the constrained domination concept. The constrained domination concept determines which solution is better than the other in multi-objective problem. Let x and y be two solutions. If both of them is infeasible, the one that violates less constraints dominates the other. If only one of them is feasible and the other is infeasible, the feasible one dominates the other. If both of them is feasible, the objective values are compared. If x is not worse than y in every objective and is strictly better than y in at least one objective, x dominates y. Applying above process to every pair of solutions in the population, each solution in the population can be ranked by the number of solutions it dominates. The objective of NSGA-II is to drive every solution towards the first rank through evolutionary process.

The NSGA-II algorithm is composed of the following steps:

1. Combine parents and offsprings in the previous generation.
2. Perform constrained non-dominated sorting within the combined population. This step produces several layers of populations with different ranking.
3. Generate the parent population for the next generation by selecting solutions from each layers in the previous step. The higher the rank of the layer, the more solutions are selected from that layer. When selecting solution in each layer, we used tournament selection.
4. Generate the offspring population for the next generation from the parent population in the previous step using genetic operations such as crossover and mutation.
5. Repeat $1 \sim 4$ for fixed number of generations.

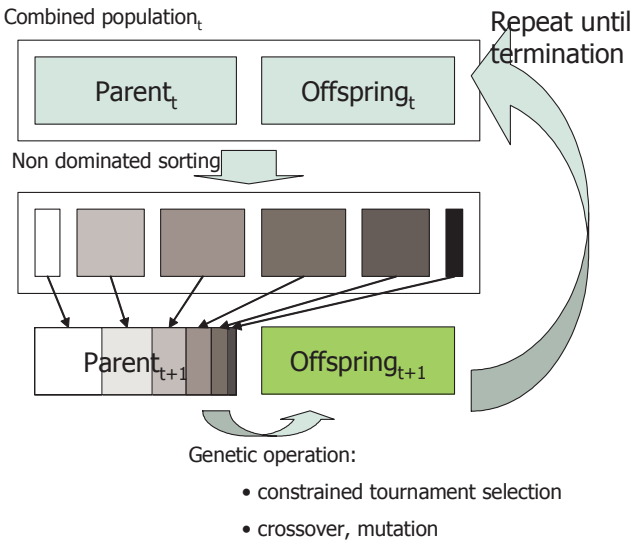

Fig. 1. The flow of NSGA-II.

The above procedure is shown in Fig. 1. While repeating above steps, the solution that has the lowest rank is removed from the population and the whole population moves towards trade-off surface. In the mean time, various trade-off solutions are evolved in the population. As a result, we can get a population of non-dominated trade-off solutions.

Among the previously mentioned conditions for good probes, we used the first condition as a constraint and the others as objectives in NSGA-II.

After the final generation, NSGA-II produces many non-dominated solutions. Among these the fittest one is chosen by using thermodynamic criteria.

3.2 The Thermodynamic Criteria

We used the thermodynamic criteria to determine if a probe hybridize to the wrong target gene. The thermodynamic criteria is based on the nearest neighbor model of DNA [8]. A probe candidate can hybridize to various site of a gene. Using the nearest neighbor model, we can calculate the free energy it takes to hybridize at each site and determine the corresponding melting temperature. If the melting temperature at a wrong hybridization site is higher than the actual hybridization temperature, misleading hybridization may occur. Using this way, we can determine if a candidate probe hybridizes to the wrong target gene or not.

Using this criteria, we chose the set of probes which have the least mis-hybridizing probes from the Pareto optimal solutions generated by multi-objective evolutionary algorithm.

4 Experimental Results

The multi-objective evolutionary algorithm introduced in the previous section is applied to design the set of probes for HPV (human papillomavirus) detection which is known to be the cause of cervical cancer. We present the result set of probes from our approach and compare its quality with those on pre-existing oligonucleotide microarray for HPV detection (Biomedlab Co., Korea) [6] in terms of specificity of hybridization.

HPV types can be divided into two classes. The HPV types which belong to one class are very likely to cause the cervical cancer and those that belong to the other class are not. 19 genotypes of HPV belong to the first class are selected as target genes. Each type of HPV has similar but different gene sequence. So our goal is to discriminate each of 19 genotypes among themselves.

The length of each probe was set to 30 nucleotides long. And based on the experimental data from Biomedlab, the hybridization temperature was set to 40°C. The concentration of sodium ion and oligomers were set to $1M$ and $1nM$ respectively.

The size of population was set as 1000 and the maximum generation number as 200. The generation number is chosen to be big enough for the population to converge to Pareto front. The crossover and mutation probabilities are set as 0.9 and 0.01 respectively.

To compare the quality of probes, we checked the specificity of hybridization for each probes. To do so, we used the following procedure. First, we compute the melting temperature of most stable configuration for every pair of probes and genes. Then, the average of the difference between these values and the melting temperature of probes is calculated. This procedure can be formulated as follows:

$$\frac{\sum_i (Tm(x_i) - \max_{j,i \neq j}(Tm(x_i, t_j)))}{|T|}$$

where, $Tm(x)$ denotes the melting temperature of probe x and $Tm(x, y)$ the melting temperature of most stable configuration of probe x and gene y. If a probe is not hybridized to its target gene specifically and is highly likely to cross-hybridize to other target genes, its melting temperature of most stable configuration with that gene would be high. Then, the difference between its own melting temperature and the cross-hybridization temperature would be small. By averaging these values for all probes, we can check specificity of hybridization of a probe set. The larger the average value is, the more specific the hybridization reactions between probes and its target genes are.

The probe set produced by the suggested algorithm has the value of 61.21 and that from Biomedlab Co. has 56.38 (see Table 1). As can be seen from the table, the suggested method could produce a more reliable probe set. The resulting probe set generated by the proposed approach is shown in Table 2.

Table 1. The comparison result between probes in commercial chip and produced sets of probes.

The Multi-objective Approach	Biomedlab. Probes
61.21	56.38

Table 2. The resulting sets of probes.

HPV Type	Probe Sequence
HPV6	GCATCCGTAACTACATCTTCCACATACACC
HPV11	GACACTATGTGCATCTGTGTCTAAATCTGC
HPV16	ACTAACTTTAAGGAGTACCTACGACATGGG
HPV18	ATGATGCTACCAAATTTAAGCAGTATAGCA
HPV31	TTGTGCTGCAATTGCAAACAGTGATACTAC
HPV33	AACTAGTGACAGTACATATAAAAATGAAAA
HPV34	GCACAAACTTTTCAGTTTGTGTAGGTACAC
HPV35	GTCTGTGTGTTCTGCTGTGTCTTCTAGTGA
HPV39	CCGTAGTACCAACTTTACATTATCTACCTC
HPV40	AGTAATTTCAAGGAATATTTGCGTCATGGG
HPV42	CAACATCTGGTGATACATATACAGCTGCTA
HPV44	CACAGTCCCCTCCGTCTACATATACTAGTG
HPV45	CCTCTACACAAAATCCTGTGCCAAGTACAT
HPV51	GGTTTCCCCAACATTTACTCCAAGTAACTT
HPV52	AGGTTAAAAAGGAAAGCACATATAAAAATG
HPV56	ACTATTAGTACTGCTACAGAACAGTTAAGT
HPV58	GCACTAATATGACATTATGCACTGAAGTAA
HPV59	ATTCCTAATGTATACACACCTACCAGTTTT
HPV66	ATTAATGCAGCTAAAAGCACATTAACTAAA

5 Conclusions

We formulated the probe design problem as a constrained multi-objective optimization problem and presented a multi-objective evolutionary method for the problem. Because our method is based on multi-objective evolutionary algorithm, it has the advantage to provide multiple choices to users. And to make it easy to choose among candidates, we suggested the thermodynamic criteria as an assistant to the decision maker. It is shown that the proposed method can be useful to design good probes by applying it to real-world problem and comparing them to currently used probes.

But in the proposed approach, several time consuming stages are contained such as the non-dominated sorting procedure. Thus, it is necessary to optimize such procedures. And our conditions for good probes are simplified and it needs more consideration on more appropriate conditions.

Acknowledgements

This research was supported in part by the Ministry of Education & Human Resources Development under the BK21-IT Program, the Ministry of Commerce,

Industry and Energy through MEC project, and the NRL Program from Korean Ministry of Science and Technology. The RIACT at Seoul National University provides research facilities for this study. The target gene and probe sequences are supplied by Biomedlab Co., Korea.

References

1. Bourneman, J. Chrobak, M., Vedova, G. D., Figueroa, A., and Jiang, T., Probe Selection Algorithms with Applications in the Analysis of Microbial Communities, *Bioinformatics*, **17**(Suppl.1):39–48, 2001.
2. Deb, K. and Goel, T., Controlled Elitist Non-dominated Sorting Genetic Algorithms for Better Convergence, *Proceedings of the First International Conference on Evolutionary Multi-Criterion Optimization (EMO-2001) LNCS 1993*, 67–81, 2001.
3. Deb, K., *Multi-Objective Optimization using Evolutionary Algorithms*, John Wiley & Sons, Ltd., England, 2001.
4. Drmanac, S., Stravropoulos, N. A., Labat, I., Vonau, J., Hauser, B., Soares, M.B., and Drmanac, R., Gene Representing cDNA Clusters Defined by Hybridization of 57,419 Clones from Infant Brain Libraries with Short Oligonucleotide Probes, *Genomics*, **37**:29–40, 1996.
5. Herwig, R., Schmitt, A. O., Steinfath, M., O'Brien, J., Seidel, H., Meier-Ewert, S., Lehrach, H., and Radelof, U., Information Theoretical Probe Selection for Hybridisation Experiments, *Bioinformatics*, **16**(10):890–898, 2000.
6. Hwang, T. S., Jeong, J. K., Park, M., Han, H. S., Choi, H. K., and Park, T. S., Detection and Typing of HPV Genotypes in Various Cervical Lesions by HPV Oligonucleotide Microarray, *Gynecol Oncol.*, **90**(1):51-56, 2003
7. Li, F. and Stormo, G.D., Selection of Optimal DNA Oligos for Gene Expression Arrays, *Bioinformatics*, **17**:1067–1076, 2001.
8. SantaLucia, John, Jr., A Unified View of Polymer, Dumbbell, and Oligonucleotide DNA Nearest-neighbor Thermodynamics, *Proceedings of National Academy of Science*, **95**:1460–1465, 1998.
9. Tobler, J. B., Molla, M. N., Nuwaysir, E. F., Green, R. D., and Shavlik, J. W., Evaluating Machine Learning Approaches for Aiding Probe Selection for Gene-expression Arrays, *Bioinformatics*, **18**(Suppl.1):164–171, 2002.