Microarray Probe Design Using
ε-Multi-Objective Evolutionary Algorithms with
Thermodynamic Criteria

Soo-Yong Shin, In-Hee Lee, and Byoung-Tak Zhang

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

Abstract. As DNA microarrays have been widely used for gene expression profiling and other fields, the importance of reliable probe design for microarray has been highlighted. First, the probe design for DNA microarray was formulated as a constrained multi-objective optimization task by investigating the characteristics of probe design. Then the probe set for human paillomavrius (HPV) was found using ε-multi-objective evolutionary algorithm with thermodynamic fitness calculation. The evolutionary optimization of probe set showed better results than the commercial microarray probe set made by Biomedlab Co. Korea.

1 Introduction

DNA microarray, especially oligonucleotide array, consists of the DNA sequences called probes, which are DNA complementaries to the genes of interest, on a solid surface. When the molecules of a cell is put to the microarray, if there exists a complementary oligonucleotide to one of the probes, it would hybridize to the probe so that a user can detect it using various methods. In this way, DNA microarray can provide the information on whether a gene is expressed or not for hundreds of genes simultaneously. Therefore, DNA microarray is widely used to study cell cycle, gene expression profiling, and other DNA-related phenomena in a cell; and has become the method of choice to monitor the expression level of a large number of genes.

By the way, microarray depends on the quality of probe sets that used. If a probe hybridizes to not only its target gene but also other genes, the microarray may produce misleading data. Thus, one needs to design the probe set carefully to get precise data. Till now, lots of probe design methods and strategies are suggested reflecting its importance [16]. Gordon and Sensen proposed a Osprey system based on various well-defined criteria [5]. Zuker group implemented OlgioArray 2.0 using thermodynamic data to predict secondary structures and to calculate the specificity of targets on chips [10]. Wang and Seed suggested OligoPicker which uses BLAST search for sequence specificity decision [18].

Though they have shown the good results, the main algorithm of most previous system is a simple *generate and filter-out approach.* Recently, a method based on machine learning algorithms such as naïve Bayes, decision trees, and neural networks has been proposed for aiding probe selection [15]. And in our previous work [8], we used a multi-objective evolutionary algorithm for probe selection of DNA microarray. We designed 19 probes for human papillomaviruses using non-dominated sorting genetic algorithm-II (NSGA-II). In this paper, we improved our previous approach in many ways. First, we reformulated the probe design problem by investigating the characteristics of the probe design. Second, we adopted ϵ-multi-objective evolutionary algorithm (ϵ-MOEA) instead of NSGA-II. In a related field, DNA sequence design for DNA computing, we noticed that ϵ-MOEA outperforms NSGA-II for DNA sequence design problem [12]. Based on these results, we improved the main algorithms to ϵ-MOEA. Third, we changed the fitness criteria of probe design by combining thermodynamic data and sequence similarity search.

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 formulate the probe design problem as multi-objective optimization problem. Section 3 and 4 describe our probe design method and provide the experimental results. In Section 5, the conclusion will be followed.

2 Multi-Objective Probe Design

2.1 Multi-Objective Optimization Problem

A multi-objective optimization problem (MOP) has a number of conflicting objectives which are to be optimized [1]. 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. However, in MOP, objectives conflict each other and there is no given priority between objectives, which makes the optimization more difficult than in single objective case.

The general form of multi-objective optimization problem is like the following:

$$
\begin{aligned}
\text{Optimize} \qquad & f_m(\mathbf{X}), \qquad m = 1, \cdots, M, \\
\text{subject to} \qquad & g_j(\mathbf{X}) \geq 0, \qquad j = 1, \cdots, N, \\
& x_i^{(L)} \leq x_i \leq x_i^{(U)}, \, i = 1, \cdots, n
\end{aligned}
\tag{1}
$$

where, $\mathbf{X}$ is a vector of n decision variable $[x_1, \cdots, x_n]^T$, f represents objective, g is constraint, M denotes the number of objectives, and N the number of constraints. $x^{(L)}$ is lower value of decision variable and $x^{(U)}$ is upper value of decision variable.

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 in the case of maximization when the following two conditions are satisfied and denoted by $X \preceq Y$:

$$\forall i \in \{i, \cdots, M\}, \ f_i((X)) \geq f_i((Y)),$$
$$\exists i \in \{i, \cdots, M\}, \ f_i((X)) > f_i((Y)). \tag{2}$$

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.

2.2   Probe Design as Multi-Objective Optimization

There exist several criteria to evaluate the set of probes [5]. We list the generally used conditions for good probes:

1. The probe sequence for each gene should not appear other genes except its target gene.
2. The probe sequence for each gene should be different from each other as much as possible.
3. The non-specific interaction between probe and target should be minimized.
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. Lastly, the probes on a oligonucleotide chip are exposed to the same experimental condition. If the melting temperatures of the probes are not uniform, some probes can not hybridize with its target.

We formulated the above conditions for clear definition of microarray probe design problem. The first condition regarded as a constraint, since it is the basic requirement for probes. And the fifth condition was not considered as one of objectives but was used as the final decision criterion to choose the best solution among diverse Pareto optimal solutions which are the results of the MOEA run.

Therefore, we formulated the microarray probe design using three fitness functions and one constraint. Before going on the formulation of the problem, let us introduce the basic notations. We denote a set of n probes by $P = \{p_1, p_2, \cdots, p_n\}$, where $p_i = \{A, C, G, T\}^l$ for $i = 1, 2, \cdots, n$, l is the length of each probe. And we denote the set of target genes by $T = \{t_1, \cdots, t_n\}$.

The constraint is the basic requirement for probes.

$$g(P) = \sum_{i \neq j} subseq(p_i, t_j), \tag{3}$$

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

Since the probe sequences should not be the subsequence of the non-target gene sequences (condition 1), this constraint is the basic requirement. And from its definition, this constraint should be zero. Other conditions are implemented as three fitness functions. First one is to prevent hybridization between probe and non-target genes (condition 2). Second is to prevent hybridization between probe and improper position of target genes (condition 3). Even though probe hybridized to the undesired site of its target gene, this can give the right information. Therefore, this seems to be unnecessary fitness functions. However, for more specific probe design, we add this fitness function in our design criteria. Last one is to prohibit forming unwanted secondary structures which can disturb the hybridization between probe and target (condition 4). They could be abstracted as follows:

$$f_1(P) = \sum_{i \neq j} hybridize(p_i, t_j), \tag{4}$$

$$f_2(P) = \sum_{i} hybridize_{target}(p_i, t_i), \tag{5}$$

$$f_3(P) = \sum_{i} secondary(p_i). \tag{6}$$

where, $hybridize(p_i, t_j)$ has non-zero value in proportion to the hybridization likelihood between p_i and t_j. $hybridize_{target}(p_i, t_i)$ is similar with $hybridize(p_i, t_j)$. It increases its value when p_i and t_i hybridize in the non-designed positions which are not the chosen site for p_i in t_i. $secondary(p_i)$ has non-zero value in accordance with the probability that p_i can form the unwanted secondary structures.
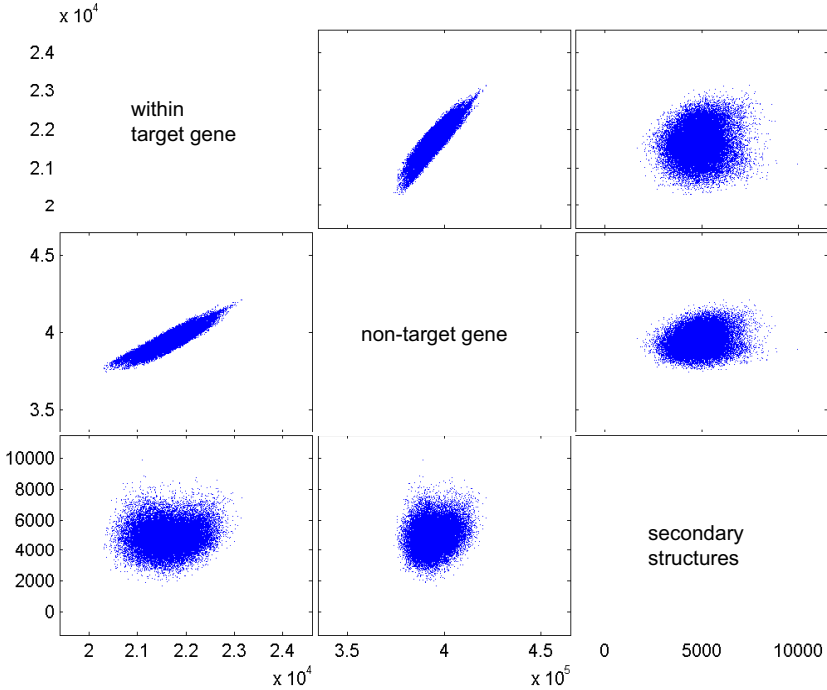
The relationship between three objectives are shown in Fig. 1. The graphs were plotted using 4^{20} 20-mer DNA sequences and their Watson-Crick complementary combinations. f_1 or f_2 has the some conflict relation with f_3. Though, in precise, the relation should be treated as random, these objectives could be solved by MOEAs. And f_1 and f_2 has a linear relation as we expected.

From above, the probe design problem is formulated as an MOP with 3 minimization objectives and 1 equality constraints.

$$\begin{aligned} \text{Minimize} \quad & f_i(P), \, i = 1, 2, 3; \\ \text{subject to} \quad & g(P) \; = 0. \end{aligned} \tag{7}$$

3   Multi-Objective Evolutionary Probe Optimization

To design probe set that satisfies above condition, we used ϵ-multi-objective evolutionary algorithm (ϵ-MOEA). There exist several methods to find such non-dominated set of solutions for a MOP. Among them, evolutionary method is one of the most popular and actively studied methods. It has the advantage that it can provide a set of non-dominated solutions by one run due to a population-based method [1]. And among various multi-objective evolutionary algorithms, ϵ-MOEA has shown the best performance [7, 2, 3].

Fig. 1. The relationship between objectives for probe design. The data are generated using 20 mer DNA sequences and Watson-Crick complement.

3.1 ϵ-Multi-Objective Evolutionary Algorithm

ϵ-multi-objective evolutionary algorithm (ϵ-MOEA) is a steady-state genetic algorithm using elite archive and ϵ-dominance relation [7, 3]. The most important characteristic of ϵ-MOEA is the ϵ-dominance relation. In ϵ-dominance relation, x ϵ-dominates y if the difference between x and y is greater than or equal to a certain amount ϵ in all objectives and is strictly better than y by ϵ in at least one objective. The mathematical definition is

$$X \quad \epsilon - dominates \quad Y \quad \Longleftrightarrow \quad (1+\epsilon)f(X) \geq f(Y). \qquad (8)$$

The ϵ-dominance is introduced to maintain a representative subset of non-dominated individuals. The ϵ-non-dominated set is smaller than the usual non-dominated set, for the non-dominated solutions which can be ϵ-dominated by others are removed in ϵ-non-dominated set. Therefore, ϵ-Pareto set is a subset of the Pareto-optimal set which ϵ-dominates all Pareto-optimal solutions. And the minimum distance between nearest solutions can be guaranteed by dividing whole search space into many grides. The density of the approximate set can be adjusted by controlling the value of ϵ [7]. Utilizing the ϵ-dominance in selecting representative subset of non-dominated set and maintaining them in the archive

1. Randomly generate initial pool.
2. Sort by domination, and set first front as archive.
3. Generate one new individual by choosing the parents from population and archive.
   (a) Choose two individuals from population.
   (b) Choose dominating solution, if dominates; choose random one, otherwise.
   (c) Choose one individuals from archive.
   (d) Perform crossover and mutation.
4. Update archive.
   (a) Replace ϵ-dominated individual(s) in the archive with new individual, if new individual ϵ-dominates archive member(s).
   (b) Leave dominating member, if there are more than one archive members in the same grid.
   (c) Add new individual, if archive members do not dominate new individual.
5. Update population.
   (a) Replace dominating individual(s) with new individual.
   (b) Replace randomly selected population member with new individual, if there is no population member which dominates the new individual.
6. Check termination.

Fig. 2. The psedocode of ϵ-MOEA

throughout generations, ϵ-MOEA showed good convergence and diversity performance [2, 3, 7].

The procedure of ϵ-MOEA for probe optimization is explained in Fig. 2. We slightly modified ϵ-MOEA proposed by Deb [2]. At each generation, parents for new offspring are chosen from the population and the archive respectively. The parent from the population is chosen by tournament selection and the parent from the archive is selected randomly. Then, an offspring is produced from these parents and evaluated. The offspring replaces an individual of the population if there exists one dominated by it in usual sense. If the offspring ϵ-dominates one or more members of the archive, it replaces the ϵ-dominated members. Or, the offspring is added to the archive if no archive member ϵ-dominates it and it ϵ-dominates no archive member. Otherwise, the offspring is discarded. Therefore, the ϵ-non-dominated individuals are always the member of the archive. This process is repeated until termination [7].

3.2 Thermodynamic Fitness Calculation

The previous microarray probe design tools can be classified into two groups by their probe specificity evaluation methods: thermodynamic approach [10, 9] and sequence similarity search approach [18, 4]. In thermodynamic approach, 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. A sequence similarity search methods used BLAST or BLAT [6] to check cross-hybridization. Since thermodynamic approach is more accurate method between them [10, 9], we calculate the fitness objectives in 2.2 using thermodynamic data. The thermodynamic fitness functions are implemented by the modified Mfold
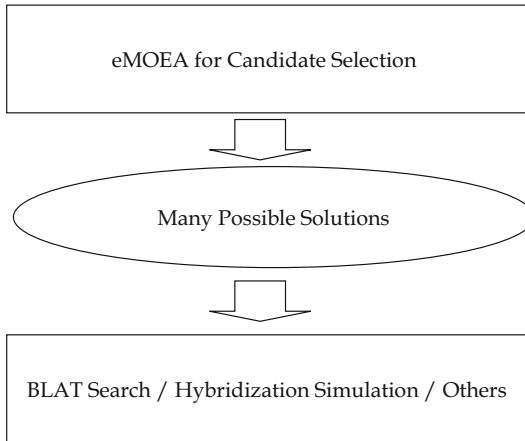
Fig. 3. The steps for probe design

[19] for OligoArray problem [10]. We downloaded the stand-alone program source code and slightly modified for fitness functions.

3.3   Probe Selection Procedure

The multi-objective evolutionary algorithm has the advantage that one can get the Pareto optimal solutions at a time. However the users usually need one promising solution, not the set of whole Pareto optimal solutions. Therefore, we incorporated the decision makers to select the most promising solution among Pareto solutions. First, the Pareto optimal solutions can be found by ϵ-MOEA. Then, BLAT search [6], hybridization simulation [13], and melting temperature calculation choose one candidate solution. BLAT is a BLAST-like sequence alignment tool, but much faster than BLAST [6]. NACST/Sim [13] is a hybridization simulation tool to check cross-hybridization based on nearest neighbor model of DNA [11]. Melting temperature is also calculated by nearest neighbor model.

Through these steps, user could be recommended the most promising probe set while maintaining the flexibility to select among various solutions. Using the characteristics of MOEA, we can improve the reliability of the optimized probe set by combining the diverse criteria such thermodynamic fitness calculation, sequence similarity search, and other user-define criteriaThis procedure is summarized in Fig. 3.

4   Experimental Results

4.1   Human Papillomavirus

The proposed constrained multi-objective approach was used to find probe set of human papillomavirus (HPV). HPV is known to be the cause of cervical cancer

[17]. HPV types can be divided into two classes: ones that are very likely to cause the cervical cancer and the others that are not. 19 genotypes of HPV belong to the first class are selected as target genes. The goal is to discriminate each of 19 genotypes among themselves. The selected 19 genes are HPV6, HPV11, HPV16, HPV18, HPV31, HPV33, HPV34, HPV35, HPV39, HPV40, HPV42, HPV44, HPV45, HPV51, HPV52, HPV56, HPV58, HPV59, and HPV66. And to improve the accuracy, L1 region of each gene sequences is chosen. Each gene and L1 region are selected by Biomedlab Co., Korea with experts' laborious works.

4.2   Parameter Settings

Based on the experimental data from Biomedlab, the length of each probe was set to 30 nucleotides long. For ϵ-MOEA, we used the various parameters. The size of population was set as 100 and the maximum generation number as 1, 1,000, 5,000, and 100,000. The crossover and mutation rates were set as 0.9 and 0.01 respectively. The ϵ was set as 1 for better convergence. For BLAT, we use default parameter settings. For NACST/Sim, we set hybridization temperature as 40°C, where the concentration of sodium ion and oligomers were set to $1M$ and $1\mu M$ respectively. The hybridization temperature was decided based on the experimental data from Biomedlab.

4.3   Probe Design Results

Our method is based on evolutionary approach, not a simple generate-filter approach which is used by most previous probe design tools. To check the merits of evolutionary approach, we compared the results by varying maximum generation from 1 to 100,000. Evolutionary algorithm with generation 1 would be the same as generate-filter method. The comparison results are shown in Table 1. As we expected, design with more generation can find better probe set. In the aspect of the number of average cross-hybridization which is checked by NACST/Sim, probe set with more generation produces the less cross-hybridization. A cross-hybridization means the undesirable hybridization between probes and genes. Especially, the comparison result between generation 1 and 1,000 showed the remarkable improvement. This means evolutionary approach can design more reliable probe set compared to the simple method. In addition, more than 1,000 generation did not show the impressive improvement. This result implies one does not need a quite large number of generation to find better probe set.

Table 1. The comparison result for various generation. As generation goes on, the probes show the less cross-hybridizations.

Generation	1	1,000	5,000	100,000
Number of average cross-hybridization	41.33	13.45	13	10.64
Number of Pareto-optimal probe sets	12	11	4	38

Table 2. *in silico* hybridization results for Pareto set with generation 1000

Set	Number of cross-hybridization
0	16
1	15
2	15
3	12
4	23
5	11
6	9
7	10
8	7
9	11
10	19

Table 3. The comparison result between probes in commercial chip (Biomedlab), selected probes using NSGA-II [8], and selected probes with ϵ-MOEA. First row means the undesirable hybridization between probes and genes calculated by NACST/Sim. Second raw represent the similar sequences appear in the wrong position. Therefore, 0 means the probe sequence appear only in its original position. The proposed method (ϵ-MOEA showed best performance for all aspects.

	ϵ-MOEA	NSGA-II	Biomedlab. Probes
# cross-hybridization	7	21	17
BLAT search	0	0 (1 for whole)	0
Melting temperature (^{o}C)	72.58 ± 3.55	74.87 ± 2.34	77.52 ± 5.03

As shown in Table 1, there are various candidate probe sets ($4 \sim 38$) as results of ϵ-MOEA. To choose best probe set among candidate probe sets, we used BLAT with HPV gene sequences first. However, we could not find any cross-hybridization using BLAT unfortunately. Since L1 region of HPV sequences is very well discriminated parts of HPV sequences, there is no similar sequences. Even when we compared L1 region sequences with whole HPV sequences using BLAT, we can find only few similar sequences. Second, we use *in silico* hybridization using NACST/Sim. The results are shown in Table 2. We used NACST/Sim for Pareto set found by 1000 generation. As explained previously, 1000 generation showed the most significant result and other runs required too much run times. As a result, we chose set no. 8 for final probe set, since that set showed the smallest number of cross-hybridization.

To verify the reliability of final probe set, we compared the probe set by ϵ-MOEA with the probes in commercial chip made by Biomedlab and the probe set by NSGA-II [8]. Table 3 showed the comparison results. ϵ-MOEA found the best probe set. Probe set by NSGA-II has three times more cross-hybridizations and Biomedlab probes has 2.5 times more cross-hybridizations. We ran BLAT for L1 region and whole HPV sequence respectively. BLAT found one similar sequences for whole HPV sequences in NSGA-II probe set, and could not find any more

Table 4. The final set of probes chosen by the proposed method for HPV

HPV Type	Probe Sequence
HPV6	CATGTACTCTTTATAATCAGAATTGGTGTA
HPV11	TCTGAATTAGTGTATGTAGCAGATTTAGAC
HPV16	TCCTTAAAGTTAGTATTTTTATATGTAGTT
HPV18	ATGTCTGCTATACTGCTTAAATTTGGTAGC
HPV31	CTTAAATACTCTTTAAAATTACTACTTTTA
HPV33	CTGTCACTAGTTACTTGTGTGCATAAAGTC
HPV34	GTGCAGTTGTACTTGTGGATTGTGTACCTA
HPV35	TTTATATGTACTGTCACTAGAAGACACAGC
HPV39	AAGGTATGGAAGACTCTATAGAGGTAGATA
HPV40	CTTGAAATTACTGTTATTATATGGGGTTGG
HPV42	AAATTAGCAGCTGTATATGTATCACCAGAT
HPV44	TTGCTTATATTGTTCACTAGTATATGTAGA
HPV45	CATGTCTACTATACTGCTTAAACTTAGTAG
HPV51	AAAGTTACTTGGAGTAAATGTTGGGGAAAC
HPV52	TTTATATGTGCTTTCCTTTTTAACCTCAGC
HPV56	ATTAATTTTTCGTGCATCATATTTACTTAA
HPV58	TTTATATGTACCTTCCTTAGTTACTTCAGT
HPV59	TAGGTGTGTATACATTAGGAATAGAAGAAG
HPV66	GAAGGTATTGATTGATTTCACGGGCATCAT

similar sequences. Probe set by ϵ-MOEA has also the lowest melting temperature among three probe sets. Though NSGA-II has the smallest melting temperature variation, the difference is not so significant compared to ϵ-MOEA. The reason why NSGA-II found the near uniform melting temperature probe set is NSGA-II used the melting temperature variation as one of objectives [8]. Even though we did not use that objective, our approach can find the comparable results. The probes practically used in Biomedlab showed the poorest results in melting temperature, even though the melting temperature variation is important for the microarray experiment protocols. The final probe set generated by the proposed approach is shown in Table 4.

5   Conclusion

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 criteria as an assistant to the decision maker. It is shown that the proposed method could be useful to design good probes by applying it to real-world problem and comparing them to currently used probes.

Though the previous works focused on finding the moderate probe set in short time, we focused on improving the quality of probe set. Therefore, our approach

need more computational time compared to the previous approaches. However, we showed the small iterations can improve the probe set quality significantly. In addition, MOEA can combine thermodynamic methods and sequence similarity search. Since these results are the preliminary results, it is necessary to optimize several time consuming stages.

Acknowledgements

References

1. K. Deb. *Multi-Objective Optimization using Evolutionary Algorithms*. John Wiley & Sons, Ltd., 2001.
2. K. Deb, M. Mohan, and S. Mishra. A fast multi-objective evolutionary algorithm for finding well-spread pareto-optimal solutions. KanGAL Report 2003002, Kanpur Genetic Algorithm Laboratory, Indian Institute of Technology Kanpur, 2003.
3. K. Deb, M. Mohan, and S. Mishra. Towards a quick computation of well-spread Pareto-optimal solutions. In *Proceedings of the Second International Conference on Evolutionary Multi-Criterion Optimization*, pages 222–236, 2003.
4. K. Flikka, F. Yadetie, A. Laegreid, and I. Jonassen. Xhm: A system for detection of potential cross hybridizations in dna microarrays. *BMC Bioinformatics*, 5(117), 2004.
5. P. M. K. Gordon and C. W. Sensen. Osprey: a comprehensive tool employing novel methods for design of oligonecleotides for dna sequencing and microarrays. *Nucleic Acid Research*, 32(17):e133, 2004.
6. W. J. Kent. BLAT–the BLAST-like alignment tool. *Genome Research*, 12(4):656–664, 2002.
7. M. Laumanns, L. Thiele, K. Deb, and E. Zitzler. Combining convergence and diversity in evolutionary multi-objective optimizatin. *Evolutionary Computation*, 10(3):263–282, 2002.
8. I.-H. Lee, S. Kim, and B.-T. Zhang. Multi-objective evolutionary probe design based on thermodynamic criteria for HPV detection. *Lecture Notes in Computer Science*, 3157:742–750, 2004.
9. F. Li and G. D. Stormo. Selection of optimal DNA oligos for gene expression arrays. *Bioinformatics*, 17:1067–1076, 2001.
10. J.-M. Rouillard, M. Zuker, , and E. Gulari. OligoArray 2.0: design of oligonucleotide probes for DNA microarrays using a thermodynamic approach. *Nucleic Acids Research*, 31(12):3057–3062, 2003.
11. J. SantaLucia Jr. A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proceedings of the National Academy of Science of the United States of America*, 95:1460–1465, 1998.

12. S.-Y. Shin. *Multi-Objective Evolutionary Optimization of DNA Sequences for Molecular Computing*. PhD thesis, School of Computer Science and Engineering, Seoul National University, Seoul, Korea, 2005.

13. S.-Y. Shin, H.-Y. Jang, M.-H. Tak, and B.-T. Zhang. Simulation of DNA hybridization chain reaction based on thermodynamics and artificial chemistry. In *Preliminary Proceedings of 9th International Meeting on DNA Based Computer*, page 451, 2004.

14. S.-Y. Shin, I.-H. Lee, D. Kim, and B.-T. Zhang. Multi-objective evolutionary optimization of DNA sequences for reliable DNA computing. *IEEE Transactions on Evolutionary Computation*, 9(2):143–158, 2005.

15. J. B. Tobler, M. N. Molla, E. F. Nuwaysir, R. D. Green, and J. W. Shavlik. Evaluating machine learning approaches for aiding probe selection for gene-expression arrays. *Bioinformatics*, 18:164–171, 2002.

16. S. Tomiuk and K. Hofmann. Microarray probe selection strategies. *Briefings in Bioinformatics*, 2(4):329–340, 2001.

17. J. M. M. Walboomers, M. V. Jacobs, M. M. Manos, F. X. Bosch, J. A. Kummer, K. V. Shah, P. J. F. Snijders, J. Peto, C. J. L. M. Meijer, and N. Munoz. Human papillomavirus is a neccsary cause of invasive cervical cancer worldwide. *Journal of Pathology*, 189(1):12–19, 1999.

18. X. Wang and B. Seed. Selection of oligonucleotide probes for protein coding sequences. *Bioinformatics*, 19(7):796–802, 2003.

19. M. Zuker. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Research*, 31(13):3406–3415, 2003.