DNA Microarray Placement For Improved Performance And Reliability

Anurag Kumar, Minsik Cho, David Z. Pan
ECE Dept. Univ. of Texas at Austin, Austin, TX 78712
{anurag, thyeros}@cerc.utexas.edu, dpan@ece.utexas.edu

ABSTRACT

DNA Microarrays or DNA chips have become a standard method for identification of DNA sequence (genes or gene mutation), determination of abundance of genes and other genomic analysis. DNA Microarrays are composed of small DNA fragments (probes) which are manufactured using masks and photolithographic method similar to VLSI industry. However, photolithographic manufacturing technology introduces errors in the probes due to unintended illumination leading to reduced performance and reliability of the Microarray. In this paper, we present a placement method using a metric for performance of microarray. We propose a Hilbert-curve filling based probe placement which incorporates the performance metric. Experimental results show upto 57x increase in the hybridization potential (leading to higher performance and reliability) of the probes on using our method.

Categories and Subject Descriptors

B.7.2 [**Hardware, Integrated Circuit**]: Design Aids

General Terms

Algorithms, Design

Keywords

Placement, microarray

1. INTRODUCTION

DNA Microarrays have become one of the most preferred methods for a number of accelerated DNA related investigations including large scale expression analysis, single nucleotide polymorphism detection, comparative genomic hybridization, mutation detection and so on [1]. DNA Microarrays are composed of small DNA fragments (called probes) synthesized at specific locations of a solid surface. Probes have a typical length of 25-60 DNA bases (adenine (A), thymine (T), guanine (G) and cytosine (C)). DNA Microarrays are synthesized using photolithographic Very Large-Scale Immobilized Polymer Synthesis (VLSIPS) [2], which is similar to the one used in the semiconductor industry where light is selectively allowed through a mask to expose cells in the array. However, such a synthesis process is not perfect as neighboring probes can get contaminated due to diffraction, scattering and internal reflection of light. This can result in synthesis of incorrect DNA sequences in the masked sites leading to unpredictable results during usage of such Microarrays. This decreases the performance and reliability of the Microarray.

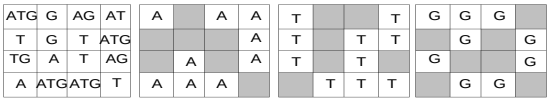
If the conditions are ideal, the functionality of DNA Microarray is not affected by the placement of probes. However, Microarrays exhibit high failure rates during manufacturing and subsequent operation. Two major sources of performance loss are (1) hybridization issues and (2) boundary issues. Hybridization issues occur because an analyte can hybridize with non-complimentary probes (which occurs with a low probability) leading to noisy results [3]. Moreover, if significant amount of hybridization does not occur between the probe and its complimentary analyte, it may not be detected by the optical detection system. Boundary issues occur at the boundary of feature of the masks during the VLSIPS process leading to wrong synthesis of probe (due to diffraction, scattering, internal reflection etc.). This results in wrong synthesis of probes. Lower hybridization potential of probes reduces their probability to bind with expected analyte and may lead to wrong detection of DNA or noisy result. Boundary issues can be reduced by better placement of probes so that the sum of border length of all masks is minimized. However, to mitigate the hybridization issues, a metric needs to be incorporated during placement to enhance the hybridization potential of the probes.

Previous DNA Microarray placement works have tried to address the boundary issues related to Microarray manufacturing rather than the hybridization issues. Boundary issues are addressed by considering probe placement a border length minimization problem. Feldman et al. first proposed an optimal solution for oligonucleotide array containing all possible probes based on two-dimensional gray code [4]. Hannenhalli et al. proposed a heuristic solution [5] by formulating it as a traveling salesman problem where each probe is a node and cost of moving from one node to the other is the hamming distance between the probes. The tour is then threaded on the two-dimensional Microarray. Border length minimization with asynchronous embedding was introduced by Kahng et. al in [6] and solved using epitaxial method. Recursive partition based algorithm was proposed in [7]. However, probe placement using border length minimization gives (1) no consideration to the thermodynamic and chemical behavior of DNA and (2) does not consider the hybridization issues related to Microarray operation. Incorporating thermodynamic stability of DNA during placement can be an effective way to characterize the hybridization potential of probes as described in Section 4. Algorithms using it [8] can be used to guide the probe placement for good microarray performance.

The major contributions of this paper include the following: We demonstrate that a simple strategy to reduce border length is not a good measure to find placement of probes on DNA Microarray and placement should be done to ensure high hybridization potential of DNA Microarrays. We propose a new cost-metric based on thermodynamic stability to account for hybridization potential of DNA Microarrays. Further, we propose a Hilbert-curve filling technique based method for thermodynamic stability based placement formulation which ensures scalability for this computationally intensive problem.

2. PRELIMINARIES

Microarrays have a regularly grided structure where each grid contains a single stranded DNA sequence called probe. Probe synthesis then starts with one nucleotide (A, T, G or C) being synthesized at each step at some selected sites. Masks are used to selectively expose the wafer to light (as shown in Fig. 1) which makes the linker molecules active at those locations, thus attaching the necleotide. The process



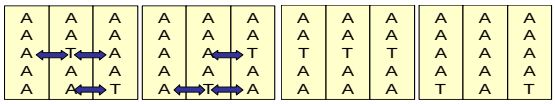
(a) (b) (c) (d)

Figure 1: Microarray shown in (a) is synthesized in three steps using three masks shown in (b), (c) and (d) respectively.

is repeated with different mask at each step to synthesize appropriate probe at the appropriate grid location. For example, to synthesize a simple Microarray shown in Fig. 1 (a), three masks are required shown in Fig. 1 (b), (c) and (d). The three masks are used in three successive steps with each mask allowing one of A, T, G or C nucleotide. In this case, mask of Fig. 1 (b) is used first to synthesize A nucleotide on the Microarray followed by masks of Fig. 1 (c) and (d) for synthesis of T and G in following steps.

3. MOTIVATION AND PROBLEM FORMULATION

DNA strands hybridize with their complimentary structure (i.e. adenine binds with thymine, guanine binds with cytosine) to form a double helical structure. However, they can also hybridize with non-complimentary structures to form secondary structures with a lower probability. Even if some errors occur during the synthesis of probes at mask borders, the resulting erroneous probe can still hybridize with the appropriate analyte particle and give expected result if the resulting structure is stable.



(a) (b) (c) (d)

Figure 2: (a) and (b) are two different placements of same probes with arrows showing the points of conflict. (c) and (d) are actually synthesized due to manufacturing errors.

Consider the problem of placing three probes of sequence length 5, AAAAA, AATAA and AAAAT on a one dimensional grid. Two possible placements having total hamming distance of 3 are shown in Fig. 2. Hence, the two placements are equivalent for border minimization method used in [5] [6]. However, consider that because of manufacturing defects, incorrect probes are synthesized(Fig. 2). We try to hybridize each resulting erroneous probe in Fig. 2 (c) and (d) with their expected complimentary strand and find the equilibrium constant using [8]. We find that while equilibrium constant for first placement (Fig. 2 (a)) is 0.82 units, for the second placement (Fig. 2 (b)) it is 1.66 units. The result shows that two different placements of probes with same hamming distance may result in markedly different performance during actual usage.

This Microarray placement problem can be formulated as follows. Let i^{th} probe p_i be of the form $n_{i1}, n_2, \cdots, n_{il}$ (where $n_{ij} \in A, T, G, C$ is a nucleotide) and $w_1, w_2, \cdots, w_l$ be the weights corresponding to different positions of the probe. Cost of placing two probes p_i and p_j at neighboring locations can be defined as:

$$C_{ij} = \sum_{k=1} w_k d(n_{ik}, n_{jk}) \quad (1)$$

where d(a,b) = 1, if a=b and d(a,b) = 0, if a≠b. The placement problem can then be defined as problem of placing the probes on the microarray such total sum of costs given by equation 1 is minimized. The mathematical formulation of the problem is given by:

$$min : \sum_{i,j} \sum_{k=1}^{l} w_k d(n_{ik}, n_{jk}) \quad (2)$$

$$subject\ to : (1)\ d(n_{ik}, n_{jk}) = 1\ \forall i, j\ \ i \neq j$$
$$(2)\ d(n_{ik}, n_{jk}) = 0\ \forall i, j\ \ i = j$$

where i,j are adjacent probes

The problem defined in Equation 2 can be seen as a three dimensional placement problem with full flexibility of moving probes in two dimensions (of Microarray) and some flexibility in third dimension (of probe [6]) by realigning the neighboring probes. Trying to solve optimally the formulation in Equation 2 would require it to be solved as a quadratic assignment problem [9]. However, this formulation is unacceptably expensive and cannot be used for a Microarray size greater than 20.

4. PERFORMANCE METRIC

A probabilistic model [3] can be used to describe the specific binding of an analyte particle with a single probe. The tendency for this reaction to reach a equilibrium is driven by the Gibb's free Energy Change ΔG^o and the relation between ΔG^0 and equilibrium constant k is defined by [10]:

$$\Delta G^o = -RT \ln k \quad (3)$$

where R is gas constant (1.987 $\frac{Calorie}{K.mole}$) and T is absolute temperature.

To analyze the effect of hamming distance on hybridization potential, we conducted an experiment by taking a 25 sequence length probe and its complimentary analyte. New probes were randomly generated which had a fixed hamming distance from original probe. Newly generated probes were hybridized with analyte and the energy released was observed. Energy released depended on the location of error in the probe. Using the effect of location of error in a probe deduced from this analysis , we propose a simple model to incorporate thermodynamic stability during placement using a weighted hamming distance where weight at i^{th} location is the equilibrium constant for hybridization of probe with its complimentary analyte with the probe having an error at i^{th} location. Algorithm 1 shows the method for finding the thermodynamic metric for placement. For each probe, we find the complimentary structure in line 3. In lines 4-8, we introduce errors at different locations of probe and hybridize it with the complimentary structure at $37^o c$. Weights are found from the energy released using Equation 3. In line 10, we normalize the weight and return it in line 11.

Algorithm 1 Thermodynamic weight characterization

Require: A set of l-length probes P
1: Array W $\Leftarrow \emptyset$
2: **for** each type $p \in P$ **do**
3: $p^{'}$ = complimentary(p)
4: **for** $i = 1$ to l **do**
5: Introduce error at i^{th} location of p
6: ΔG^o = hybridize$(p, p^{'})^{Temp.=37^o c}$
7: $W_i +=$ $exp(\frac{-\Delta G^o}{RT})$
8: **end for**
9: **end for**
10: W = $\frac{W}{count(P)}$
11: **return** W

4.1 Algorithm

We adopt a Hilbert-curve filling based strategy for placement of DNA Microarray. The key observations that we use to guide our placement are the following. The thermodynamic weights at the center is much higher than the weights

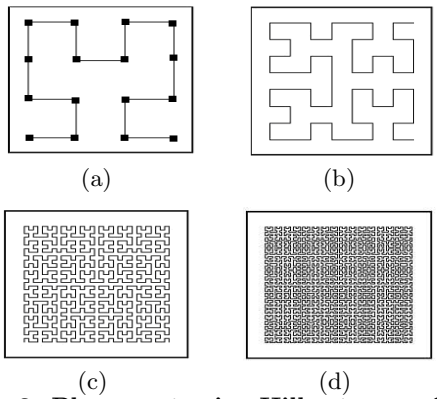


Figure 3: Placement using Hilbert curve filling technique for 4×4, 8×8, 32×32 and 64×64 size Microarray.

at the tail of the probe. Moreover, probes are generally selected in such a way that any two probe have a large difference between them [11]. This observation along with the fact that all probes are independent and uncorrelated can be used to solve the problem by partitioning. The solution can also be improved by re-embedding [6] and re-shuffling of probes. These observations lead us to adopt a Hilbert-curve filling strategy for placement. Hilbert-curve filling technique has an inherent partitioning property and maintains a low cost between the probes in two dimension.

A one-dimensional ordering of probes is first found by formulating this problem as a graph-traversal problem where each node is a probe and cost of moving from one node to other is given by equation 1. The sequence of arrays is obtained by traversing this graph such that each node is traversed once and the traversing cost is minimal. For example for a Microarray size 4×4 with 16 probes to be placed, two dimensional placement can be obtained from one dimensional ordered probes by placing the probes on the dots and moving along the line starting with bottom right corner of Fig. 3 (a). For large Microarray size, the method can be extended similar to those shown in Fig. 3 (b), (c) and (d) for Microarray size 8×8, 32×32 and 64×64.

Algorithm 2 Thermodynamic stability based placement

Require: A set of l-length probes P, A 2-D Microarray M of size s

1: Compute thermodynamic weights W
2: Partition $T = \phi$
3: **while** l>threshold **do**
4: v = partition(l)
5: $T = T \cup v$
6: **end while**
7: **for** each T **do**
8: O = 1-D ordering of T
9: M = Place O using Hilbert-curve filling
10: **end for**
11: **return** M

The overall algorithm is described in algorithm 2. First we extract the thermodynamic weights for the probes. In lines 2-6, we partition the problem into manageable size. Partitioning of the problem largely depends on the time taken in finding the one-dimensional ordering of the probes. In lines 7-10, we do the placement for each partition. For each partition, a one dimensional ordering is first found followed by placement using Hilbert-curve filling [12]. Conflicts between neighboring probes are further reduced using sequence alignment between probes [13]. The placement generated by Hilbert-space filling technique works well for the following reasons: Probes with low cost amongst themselves in one

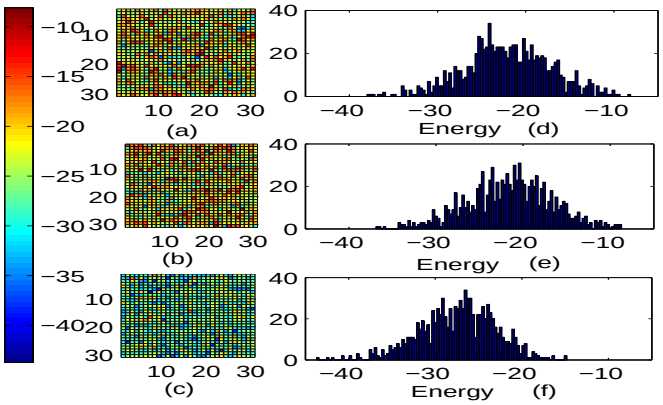


Figure 4: Energy released for different placement results of a 32×32 Microarray.

dimensional ordering remain close after the two dimensional placement. Hilbert-space filling curve has an inherent partitioning property which makes the partitioning of problem easy. Hilbert-space fills can be generated for individual partitions and then stitched together.

5. EXPERIMENTAL RESULTS

We have implemented the Hilbert-curve filling based placement described in Section 4.1 in C. For comparison, we implemented TSP+1-threading [5] and lexicographic-sorting [6] based methods. Probes were generated randomly for different Microarray size. All algorithms were tested on an Intel Xeon 2.4 GHz Linux machine with 4GB RAM.

Table 1: Total Border length for different placement results

Size	TSP+1-threading [5]	lexico [6]	Our algorithm
8	1765	1950	1598
16	7969	8533	7068
32	33272	35775	29141
64	132712	145399	116440
128	516440	575521	455673
norm.	1.00	1.11	0.88

We first compared TSP+1-threading [5] and lexicographic sorting based method with row-epitaxial [14] against Hilbert-curve filling method based placement for border costs. Table 1 shows the result of experiments performed on 25-length probes for different Microarray size generated randomly. Placement using our algorithm shows an average improvement of 12% over TSP+1-threading and 20.5% improvement over lexicographic sorting method in terms of total border length. Next, we conducted experiments to show the efficacy of our Thermodynamic-stability based placement. Since, most of the errors occur at the border of the masks, we introduce errors at such border locations randomly causing change in the probes. After introducing errors, each manipulated probe was hybridized with the actual complimentary analyte it was expected to hybridize with and the thermodynamic energy released was found at 37^oC temperature using [8].

We first show the placement result for a 32×32 Microarray in Fig. 4 (average error per probe = 6). Fig. 4 (a), (b) and (c) show the energy released for placements using TSP+1-threading, lexicographic sorting and our method with thermodynamic considerations respectively. Colors in each grid corresponds to the energy released (Cal./kmol) by the probes with blue representing -45 Cal./kmol and red representing -5 Cal./kmol. To analyze the placement result, we have shown the histogram of energy released in Fig. 4 (d),

Table 2: Comparison between different placement results for Thermodynamic-stability

Array Size	TSP+1-threading [5]			lexico [6]			Weighted+Our Algorithm			Thermo+Our Algorithm		
	Energy[a]	Avg.[b]	Hybrid.[c]	Energy	Avg.	Hybrid.	Energy	Avg.	Hybrid.	Energy	Avg.	Hybrid.
Average error per probe 6												
8	-1170	-18.28	1.70e13	-1192	-18.62	3.00e13	-1360	-21.25	2.40e15	-1379	-21.55	3.97e15
16	-5530	-21.60	4.21e15	-5467	-21.35	2.84e15	-6013	-23.49	1.00e17	-6225	-24.43	4.82e17
32	-23899	-23.24	6.63e16	-23684	-23.13	5.52e16	-27035	-26.40	1.28e19	-26966	-26.33	1.14e19
64	-99890	-24.39	4.51e17	-97382	-23.77	1.60e17	-108742	-26.54	1.62e19	-99982	-24.41	4.46e17
128	-406940	-24.84	9.54e17	-397914	-24.29	3.82e17	-441449	-26.94	3.19e19	-449516	-27.44	7.27e19
total	-537429	-	1.47e18	-525639	-	6.00e17	-584599	-	6.10e19	-584068	-	8.50e19
norm	1.0	-	**1.0**	0.97	-	**0.40**	1.08	-	**41.49**	1.09	-	**57.82**
Average error per probe 1												
8	-1869	-29.20	1.37e21	-1881	-29.39	1.88e21	-1911	-29.86	4.10e21	-1896	-29.64	2.84e21
16	-8598	-33.59	2.06e24	-8536	-33.35	1.38e24	-8552	-33.40	1.50e24	-8594	-33.57	1.99e24
32	-36763	-35.90	9.67e25	-36606	-35.75	7.72e25	-36656	-35.79	8.05e25	-37052	-36.18	1.54e26
64	-152569	-37.25	9.17e26	-151807	-37.06	6.68e26	-151897	-37.08	6.91e26	-152863	-37.25	9.17e26
128	-620199	-37.85	2.49e27	-618152	-37.73	2.04e27	-617468	-37.69	1.91e27	-624112	-38.09	3.72e27
total	-819998	-	3.50e27	-816982	-	2.78e27	-816484	-	2.68e27	-824517	-	4.79e27
norm.	1.0	-	**1.0**	0.99	-	**0.79**	0.99	-	**0.76**	1.005	-	**1.37**

[a] Energy released in calorie per kilomole
[b] Average energy released in calorie per kilomole per probe
[c] Hybridization equilibrium constant

(e) and (f) corresponding to placements shown in Fig. 4 (a), (b) and (c) respectively. Mean and standard deviation of (d), (e) and (f) is (-22.1,4.98), (-21.5,4.88) and (-27.4,4.36) Cal./kmol. respectively. Mean of energy released for Fig. 4 (f) is 22% lower than (d). Table 2 shows the overall result (with error rate per probe of 6 and 1). We have shown results of hilbert-curve filling technique with simple weighted distance as well as thermodynamic weights. Equilibrium constant reported for every placement is direct indicator of performance of the microarray for a given placement. Higher equilibrium constant shows higher chance and amount of hybridization between the probe and analyte leading to easy detection. Comparing the equilibrium constants for different placements, we find that for one error per probe, our method shows an average improvement of 1.37 times over the other methods. For a high error per probe of six, our algorithm shows 57 times better result than other method for placement. It can be noted that robustness of placement increases exponentially with increase in the number of errors present in the probe. This is because of the exponential relationship between the thermodynamic energy released during hybridization and the equilibrium constant. The results strongly suggest that making the placement of probes for Microarray performance-aware can produce significantly better results.

6. CONCLUSIONS

To increase the performance and reliability of DNA microarrays, we presented a placement method and a new equilibrium constant based metric. The experimental results show a significant improvement in hybridization potential of probes using our algorithm compared to conventional methods.

7. REFERENCES

[1] M. Schena, *Microarray Analysis*. Wiley, New York, 2003.

[2] S. Fodor, J. Read, M. C. Pirrung, L. Stryer, A. T. Lu, and D. Solas, "Light-directed, spatially addressable parallel chemical synthesis," *Science*, pp. 175–181, 1991.

[3] A. Hassibi, H. vikalo, and A. Hajimiri, "On noise processes and limits of performance in biosensors," *Journal of Applied Physics*, p. 102, 2007.

[4] W. Feldman and P. Pevzner, "Gray code masks for sequencing by hybridization," *Genomics*, vol. 23, pp. 233–235, 1994.

[5] S. Hannenhalli, E. Hubell, R. Lipshutz, and P. Pevzner, "Combinatorial algorithms for design of DNA arrays," *Advances in Biochemical Engineering Biotechnology*, pp. 1460–1465, 2002.

[6] A. B. Kahng, I. I. Mandoiu, P. A. Pevzner, S. Reda, and A. A. Zelikovsky, "Border length minimization in DNA array design," in *Proc. 2nd Workshop on Algorithms in Bioinformatics*, pp. 435–448, 2002.

[7] A. B. Kahng, I. I. Mandoiu, S. Reda, X. Xu, and A. Zelikovsky, "Evaluation of placement techniques for DNA probe array layout," in *Proc. Int. Conf. on Computer Aided Design*, pp. 262–269, 2003.

[8] N. Markham and M. Zuker, "Dinamelt web server for nucleic acid melting prediction," *Nucleic Acid Res.*, pp. 577–581, 2005.

[9] P. Pardalos and H. Wolkowicz, "The quadratic assignment problem: a survey of recent developments," *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, 1994.

[10] R. S. Berry, S. A. Rice, and J. Ross, *Physical Chemistry*. Oxford University Press, 2000.

[11] F. Li and G. Stormo, "Selecting optimum DNA oligos for microarrays," in *IEEE International Symposium on Bio-Informatics and Biomedical Engineering*, pp. 200–207, 2000.

[12] A. R. Butz, "Alternative algorithm for Hilbert's Space-Filling curve," *IEEE Transactions on Computers*, pp. 424–426, 1971.

[13] X. Huang, "On global sequence alignment," *Comput. Appl. Biosci.*, vol. 10, pp. 227–235, 2003.

[14] A. B. Kahng, I. I. Mandoiu, S. Reda, X. Xu, and A. Zelikovsky, "Design flow enhancements for DNA Arrays," in *Proc. IEEE Intl. Conf. on Computer Design*, October 2003.