Improving the Layout of Oligonucleotide Microarrays

Sérgio A. de Carvalho Jr.[1] and Sven Rahmann[2]

[1] Graduiertenkolleg Bioinformatik, Bielefeld University, D-33594 Bielefeld, Germany.
Sergio.Carvalho@cebitec.uni-bielefeld.de
[2] Algorithms and Statistics for Systems Biology group, Genome Informatics,
Faculty of Technology, Bielefeld University, D-33594 Bielefeld, Germany.
Sven.Rahmann@cebitec.uni-bielefeld.de

Abstract. The production of most commercial microarrays is based on a parallel chemical synthesis driven either by a set of masks or micromirror arrays. Unfortunately, both mechanisms can experience problems due to stray light if the arrangement of probes is not carefully designed. This paper presents means of measuring the quality of a design and examines the layout of several existing microarrays. It also describes an algorithmic strategy to minimize the risk of unintended illumination.

1 Introduction

A DNA microarray is a piece of glass or plastic on which single-stranded fragments of DNA (called *probes*) are affixed or synthesized. The chips produced by Affymetrix can contain more than one million spots, with each spot accommodating several million copies of a probe. Probes are typically 25 bases long and are synthesized in parallel, on the chip, in a series of repetitive steps. Each step appends the same nucleotide to probes positioned in selected regions of the chip. Selection occurs by exposure to light with the help of a photolithographic mask that allows or obstructs the passage of light accordingly [1].

Formally, we have a set of probe sequences $\mathcal{P} = \{p_1, p_2, ...p_n\}$ that are produced by a series of masks $\mathcal{M} = (m_1, m_2, ...m_\mu)$, where each mask m_i induces the addition of a particular nucleotide $\nu_i \in \{A, C, G, T\}$ to a subset of $\mathcal{P}$. The *nucleotide deposition sequence* $\mathcal{S} = \nu_1 \nu_2 \ldots \nu_\mu$ corresponding to the sequence of nucleotides added at each masking step is therefore a supersequence of all sequences from $\mathcal{P}$.

In general, a probe can be *embedded* within $\mathcal{S}$ in several ways. We can think of an embedding of p_j as a μ-tuple $(e_{j,1}, e_{j,2}, ...e_{j,\mu})$ in which $e_{j,i}$ equals 1 if probe p_j receives nucleotide ν_i (at step i), or 0 otherwise. We say that the embedding of p_j is *productive* at step i when $e_{j,i} = 1$, or *unproductive* when $e_{j,i} = 0$. Equivalently, given an arrangement of the probes on the chip, we can say that a spot s is productive at step i if mask μ_i allows the passage of light to s (probes of s are activated to chemical coupling with nucleotide ν_i), or unproductive otherwise.

2 Border Length

Due to diffraction of light or internal reflection, untargeted spots can sometimes be accidentally activated in a certain step, producing unpredicted probes that can compromise the results of an experiment. This issue was first addressed by Hannenhalli et al. [2]. They noted that the problem is more likely to occur near the borders between masked and unmasked spots. They were the first to formulate the Border Length Minimization Problem, which aims at finding an arrangement of probes that minimizes the number of border conflicts during mask exposure steps.

Given a mask μ_i we compute its border length as the number of borders shared by masked (unproductive) and unmasked (productive) spots. The total border length of a given arrangement is the sum of border lengths over all masks.

3 Conflict Index

In [3], Kahng et al. noted that the definition of border length did not take into account two simple yet important practical considerations: a) stray light might activate not only immediate neighbors but also probes that lie as far as three cells away from the targeted spot; and b) imperfections produced in the middle of a probe are more harmful than in its extremities. With these observations in mind, we define the conflict index $\kappa(s)$ of a spot s whose probes of length ℓ_s are synthesized in μ masking steps as follows. First we define a distance-dependent weighting function, $\delta(s, s', i)$, that accounts for observation a) above:

$$\delta(s, s', i) := \begin{cases} 0 & \text{if spot } s' \text{ is masked at step } i, \\ \frac{1}{(d(s,s'))^2} & \text{otherwise,} \end{cases} \tag{1}$$

where $d(s, s')$ is the Euclidian distance between spots s and s'. We also use position-dependent weights as suggested in [3] to account for observation b):

$$\omega(s, i) := \begin{cases} 0 & \text{if spot } s \text{ is unmasked at step } i, \\ \sqrt{\min(b_{s,i} + 1, \ell_s - b_{s,i} + 1)} & \text{otherwise,} \end{cases} \tag{2}$$

where $b_{s,i}$ denotes the number of nucleotides synthesized at spot s up to and including step i. We now define the conflict index of a spot s as

$$\kappa(s) := \sum_{i=1}^{\mu} \left(\omega(s, i) \sum_{s'} \delta(s, s', i) \right), \tag{3}$$

where s' ranges over all spots near s (in practice, only those inside a 7x7 grid centered in s).

It should be clear that our definition of conflict index is intimately connected to that of border length although the first estimates the risk of producing faulty probes in a given spot while the latter measures the quality of a mask.

4 Analyzing Affymetrix Microarrays

We examined the layout of a number of Affymetrix GeneChip® arrays with regard to border length and our definition of conflict index. With this analysis we could distinguish different placement strategies. Some of them were visibly better than others when it came to reducing spot conflict (and border length).

Figure 1 shows the normalized border length per masking step (border length divided by the number of probe sequence) for selected GeneChip arrays. Clearly, the E. Coli 1.0 chip has, overall, the highest border length per masking step. Most of the earlier chips produced by Affymetrix have analogous curves (data not shown), which suggests that this was one of their first placing strategies.

The curves of the Human U95-A and the Human U133 Plus 2.0 chips in Fig.1 are similar except for the first 13 masks. This suggests that an improvement in the placing algorithm of the latter was able to reduce the border length of its first masks.

In fact, all of the latest chips including the Human U133 Plus 2.0 have been designed with the following simple strategy. The chip is first divided into two horizontal bands. Probes whose embeddings are productive at the first masking step are assigned to the lower band, while the others are assigned to the upper band. Each band is then recursively divided into two horizontal bands and probes are assigned to them according to the state of their embeddings in the second masking step, in such a way that the productive bands of each partition are next to each other — the ordering of assignments obey a (one-dimensional) gray code (see Fig.2a). This process is repeated a few times until the bands become too narrow to be divided. The resulting masks consist of alternating bands of productive and unproductive spots, which effectively reduces their border length.

Figure 3 shows the distribution of conflict indices for three of the largest GeneChip arrays designed with the one-dimensional partitioning strategy described above. The layout of the Human Genome chip has undoubtedly reduced the amount of conflicts. The reason is the following.

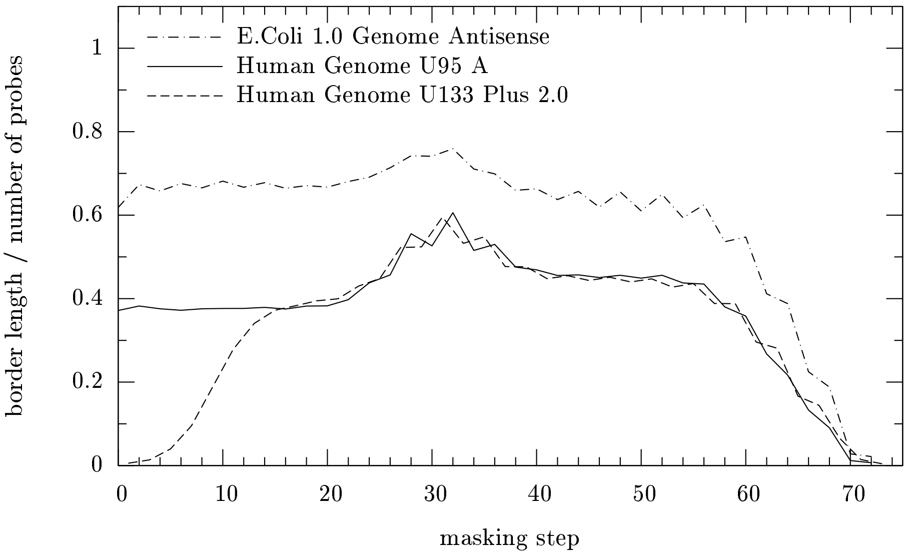

Fig. 1. Normalized border length per masking step for selected GeneChip arrays

This chip has more than 1.3 million spots which are not fully utilized. Around 10.4% of them do not contain probes and the layout is such that the empty spots are evenly distributed throughout the array. In contrast, the Chicken chip (with as much as 11.9% of empty spots) and the Rice chip (with 6.5% of empty spots) have their probes concentrated on the upper part of the array with the lower part consisting of a large empty area. Clearly, they could have benefited from having a more uniform distribution of empty spots. It is not clear why such approach have been taken. In fact, this came as a surprise given that they are among the latest GeneChip arrays available.

5 Two-Dimensional Gray Code Partitioning

It should be clear from Fig.2a that the one-dimensional partitioning employed by Affymetrix cannot optimize many masks because the regions soon become too small to be divided. We are currently investigating a similar partitioning strategy that has the potential of optimizing twice as many masks. The main differences are that our partitioning alternates between horizontal and vertical divisions, and that we assign probes to regions based on a two-dimensional gray code.

The inspiration for our approach came from Feldman and Pevzner [4] who proved that a placement of the complete set of strings of length l over a four-letter alphabet based on a two-dimensional gray code has an optimal border length. However, their work was based on the following assumptions: a) a chip of $2^l \times 2^l$ spots must have the complete set of l-tuples; b) the deposition sequence must be periodic, e.g. $\mathcal{S} = (ACGT)^k$ and c) probes must be *synchronously* embedded, i.e. probes can only receive one nucleotide of a given period of $\mathcal{S}$. In practice, these assumptions are unrealistic. The complete set of 25-base-long probes would require a chip with $2^{25} \times 2^{25}$ spots, and a synchronous embedding would increase the number of masking steps to $25 \times 4 = 100$ (most of the chips produced by Affymetrix are built in 74 steps).

Our approach, on the other hand, does not need any of the above assumptions to work and proceeds similarly to the method employed by Affymetrix. First, we partition the chip into two horizontal bands and assign probes to each of them according to whether their embeddings are productive or unproductive at the first masking step. Each resulting region is then independently partitioned into two vertical bands while probes are separated according to the state of their embeddings at the second masking step. As mentioned earlier, assignments obey a two-dimensional gray code. This procedure is recursively repeated several times, producing a set of masks as depicted in Fig.2b.

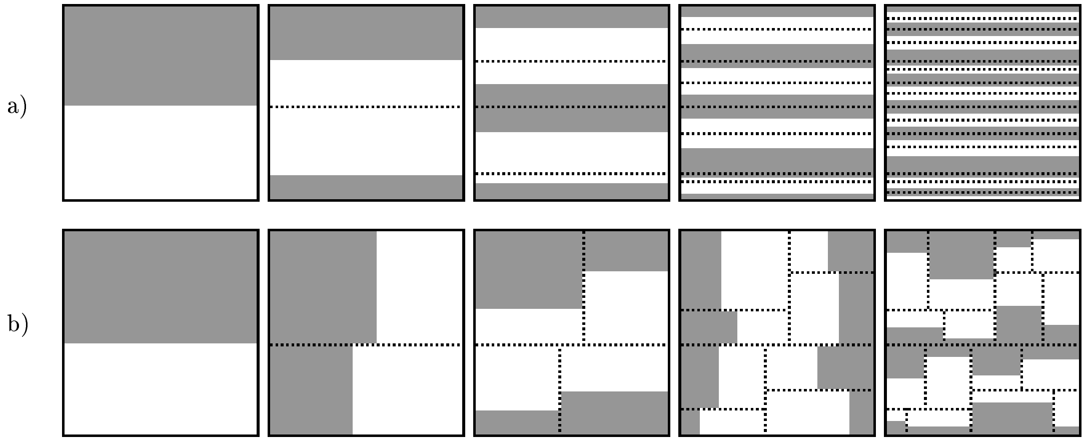

Fig. 2. Photolithographic masks resulting from two partitioning strategies. Shaded areas contain masked spots (unproductive) whereas unmasked spots are concentrated in the white regions. Dotted lines only highlight previous partitions. a) Affymetrix's one-dimensional partitioning; b) our two-dimensional partitioning

6 Results

An early implementation of our placement algorithm confirms that it can optimize twice as many masks when compared to the algorithm employed by Affymetrix. However, it is clear that such approaches cannot optimize the complete set of masks.

In fact, our first experiments show that better layouts can be produced if the divisions stop before the regions become too small. The reason is the following. While we can optimize as much as 26 of the first masking steps, the arrangement of the remaining masks is left completely random and there is little freedom for optimizing them. In contrast, if the division stops while these regions are reasonably large, it is possible to choose the best location for each probe so that the complete set of masks is optimized.

Figure 4 shows the normalized border length of the E.Coli 2.0 Genome chip as originally designed by Affymetrix compared to the layout produced by our placement algorithm in an attempt to optimize as many of the first masking steps as possible. As expected, while it was possible to significantly optimize the first masks, the border length of the remaining steps has increased.

Our goal is thus to combine our algorithm with another optimization technique that can reduce the overall border length. At the moment, we are evaluating some alternatives described recently [5] – including those that re-embed probes optimally in regard to their neighbors – so that we can fully evaluate our approach.

References

1. Fodor, S., Read, J., Pirrung, M., Stryer, L., Lu, A., Solas, D.: Light-directed, spatially addressable parallel chemical synthesis. Science **251** (1991) 767–73
2. Hannenhalli, S., Hubell, E., Lipshutz, R., Pevzner, P.: Combinatorial algorithms for design of DNA arrays. Adv. Biochem. Eng. Biotechnol. **77** (2002) 1–19
3. Kahng, A. B., Mandoiu, I., Pevzner, P., Reda, S., Zelikovsky, A.: Engineering a scalable placement heuristic for DNA probe arrays. Proceedings of the Seventh Annual International Conference on Computational Molecular Biology (2003) 148–83
4. Feldman, W., Pevzner, P.: Gray code masks for sequencing by hibridization. Genomics **23** (1994) 233–235
5. Kahng, A. B., Mandoiu, I., Reda, S., Xu, X., Zelikovsky, A.: Evaluation of placement techniques for DNA probe array layout. Proceedings of the IEEE/ACM International Conference on Computer-Aided Design (2003) 262–269

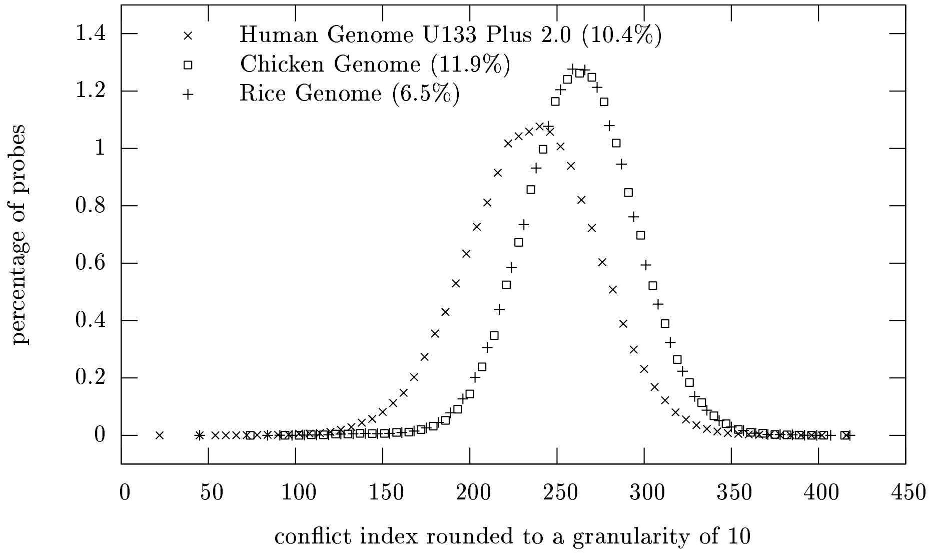

Fig. 3. Distribution of conflict index values for selected GeneChip arrays (the amount of empty spots is shown in brackets). The layout of the Human U133-P2 chip has reduced conflicts because of an even distribution of empty spots throughout the array

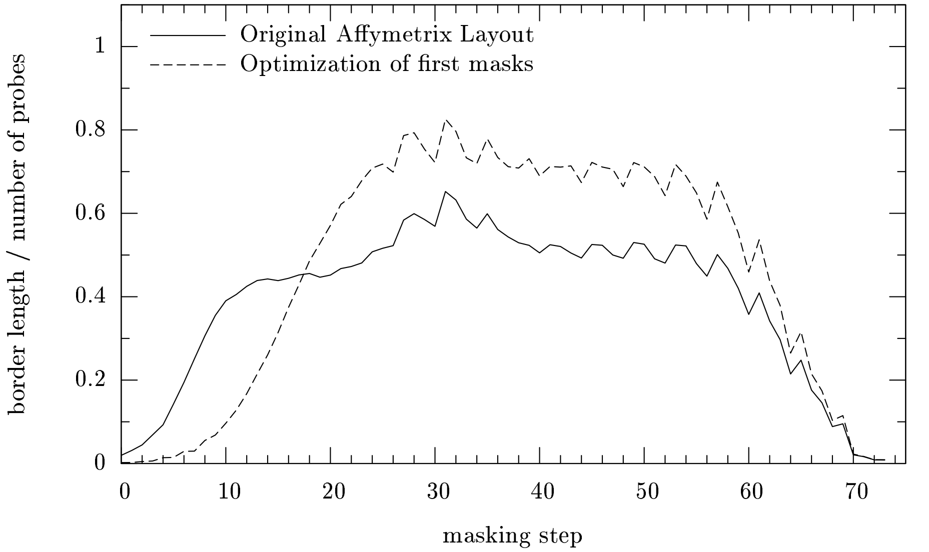

Fig. 4. Normalized border length of the E.Coli 2.0 Genome chip as originally designed by Affymetrix compared to an attempt to optimize the first masks with our two-dimensional partitioning