BIOINFORMATICS

Vol. 00 no. 00 2006
Pages 1–3

Quadratic Assignment and the Layout of Oligonucleotide Microarrays

Sérgio A. de Carvalho Jr.[a] *and Sven Rahmann*[b]

[a] *Graduiertenkolleg Bioinformatik, Bielefeld University, Germany,*
[b] *Algorithms and Statistics for Systems Biology, Genome Informatics, Bielefeld University, Germany.*

ABSTRACT

Motivation: The production of commercial DNA microarrays is based on a light-directed chemical synthesis driven by a set of masks or micromirror arrays. Due to the natural properties of light and the ever shrinking feature sizes, the arrangement of the probes on the chip and the order in which their nucleotides are synthesized play an important role on the quality of the final product. In this paper, we review existing models and algorithms for designing high-density oligonucleotide microarrays. We also define an extended model to evaluate microarray layouts and investigate a new approach based on the *quadratic assignment problem* (QAP).

Results: We used an existing QAP heuristic algorithm to design the layout of small artificial microarrays with excellent results. We compare this approach with the best known algorithm and describe how it can be combined with other existing algorithms to design the latest million-probe chips.

Availability: Source code is available from the authors upon request.

Contact: Sergio.Carvalho@cebitec.uni-bielefeld.de

1 INTRODUCTION

An oligonucleotide 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, for instance, can contain more than one million spots (or *features*) as small as 11 μm, with each spot accommodating several million copies of a probe. Probes are typically 25 nucleotides long and are synthesized in parallel, on the chip, in a series of repetitive steps. Each step appends the same nucleotide to probes of 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 (Fodor *et al*., 1991).

Formally, we have a set of probes $\mathcal{P} = \{p_1, p_2, ...p_n\}$ 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 $\nu_i \in \alpha = \{A, C, G, T\}$ 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 $p_i \in \mathcal{P}$. In general, a probe can be *embedded* within $\mathcal{S}$ in several ways. An embedding of p_j is a μ-tuple $\varepsilon = (e_{j,1}, e_{j,2}, ...e_{j,\mu})$ in which $e_{j,i} = 1$ if probe p_j receives nucleotide ν_i (at step i), or 0 otherwise (Figure 1).

Fig. 1. Synthesis of a hypothetical 3x3 chip. a) Chip layout and 3-base-long probe sequences; b) deposition sequence and embeddings of the highlighted probes; c) the first three resulting photolithographic masks.

Deposition sequences are usually cyclical, that is $\mathcal{S}$ is a repeated permutation of α. This is mainly because such sequences maximize the number of possible subsequences (Chase, 1976). In this context, we can distinguish between *synchronous* and *assynchronous* embeddings. In the first case, each probe synthesize one and only one nucleotide in every cycle of the deposition sequence. Thus, for probes of length 25, we need 100 masking steps. In the case of assynchronous embeddings, each probe can syntesize any number of nucleotides in any given cycle. This allows for shorter deposition sequences. Most (if not all) Affymetrix chips, for instance, can be synthesized in 74 masking steps.

Due to diffraction of light or internal reflection, untargeted spots can sometimes be accidentally activated in a certain masking step, producing unpredicted probes that can compromise the results of an experiment. This issue was described by Fodor *et al*., 1991 who noted that the problem is more likely to occur near the borders between masked and unmasked spots. This observation has given rise to the term *border conflict*.

We are interested in finding an arrangement of the probes on the chip together with their embeddings in such a way that we minimize the chances of unintended illumination during mask exposure steps. As we show in later sections, this problem is intrinsically hard and optimal solutions are unlikely to be found even for very small chips with fixed embeddings due to the exponential number of possible arrangements.

If we consider all valid embeddings, the problem is even harder. A typical probe of an Affymetrix chip, for instance, can have up to several million possible embeddings. For this reason, the problem has been traditionally tackled in two steps. First, an initial embedding of the probes is fixed and an arrangement of these embeddings on the chip with minimum border conflicts is sought. This is usually refered to as the *placement* problem. The second step is a *post-placement* optimization that re-embeds the probes considering its location on the chip, in such a way that the conflicts with the neighboring spots are further reduced.

In the next section, we review the Border Length Minimization Problem introduced by Hannenhalli *et al*, 2002, and define an extended model for evaluating microarray layouts. In section 3 we briefly review exising placement strategies. In section 4 we propose a new approach to the problem based on a quadratic assignment problem (QAP) formulation. We present the results of using a QAP heuritic algorithm to design small artificial chips in section 5, along with a comparison with the best known placement algorithm. In section 6 we describe how this approach can be used to design larger microarrays.

2 MODELING

Hannenhalli *et al*, 2002 were the first to give a formal definition to the problem of unintended illumination in the production of microarrays. They formulated the Border Minimization Problem, which aims at finding an arrangement of the probes together with their embeddings in such a way the number of border conflicts during mask exposure steps is minimal.

The *border length* of a mask μ_i is defined as the number of borders shared by masked and unmasked spots at masking step i. The total border length of a given arrangement is the sum of border lengths over all masks.

2.1 Conflict Index

Kahng *et al*, 2003-1 noted that the definition of border length does 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 to account for observation b):

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

where $b_{s,i}$ denotes the number of nucleotides synthesized at spot s up to and including step i.

TODO: review this definition (use of exp?).

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).

The definition of border length is somewhat connected to our definition of conflict index. However, while the first measures the quality of a mask, the latter estimates the risk of producing faulty probes in a given spot.

TODO: elaborate a bit more about the relation between border length and conflict index. Explain how the choice of position-dependent or distance weighting function can influence the border lenght when trying to reduce the conflict index. Also, mention that Kahng *et al*, 2003-1 suggested to use square root for the position-dependent and that we decided to use an exponential relation.

3 PREVIOUS WORK

In this section we review existing algorithmic techniques for designing oligonucleotide microarrays.

Feldman and Pevzner, 1994 were the first to formally address the border length problem. They described an optimal solution based on a 2-dimensional Gray code. However, their work is restricted to *uniform arrays* (arrays containing all possible probes of a given length) and synchronous embeddings.

Hannenhalli *et al*, 2002 were the first to work with arrays of arbitrary probes. They reported that the first Affymetrix arrays were designed using a heuristic algorithm for the traveling salesman prolem (TSP). The idea consisted of building a weighted graph with nodes representing probes and edges containing the hamming distance between the probes. Then, a TSP tour with minimum weight was constructed, resuling in consecutive probes in the tour being likely to be similar. The TSP tour was then *threaded* on the array in a row-by-row fashion. Hannenhalli *et al*, 2002 enhanced this approach by suggesting a different threading of the TSP tour on the chip, called *1-threading*, to achieve up to 20% reduction in border length.

Kahng *et al*, 2002 suggested an *epitaxial* placement algorithm for arrays with synchronous embeddings. Their algorithm places a random probe in the center of the array

and continues to insert probes in spots adjacent to already placed probes. It employs a greedy heuristic to select the next sequence to be placed among all non-placed probes in such a way that the number of border conflicts is reduced. With this algorithm, they claimed to achieve up to 10% reduction in conflicts over the TSP-based approach of Hannenhalli *et al*, 2002.

Although we believe that solution quality is more important than running time, the major problem with the epitaxial and the TSP-based algorithm, as noted in Kahng *et al*, 2003-1, is that they have at least quadratic time complexity and thus are not scalable for the latests million-probe microarrays.

This observation has led to the development of two new algorithms in Kahng *et al*, 2003-1. The first one is a simple variant of the epitaxial algorithm described in Kahng *et al*, 2002, called row-epitaxial. The main differences are: a) spots are filled in a pre-defined order, namely row-by-row; and b) a limited number of candidates are considered when filling each spot.

The other algorithm, called sliding-window matching (SWM), needs an initial placement that can be constructed by, for instance, TSP and 1-threading. It iteratively improves the current placement by selecting an independent set of spots inside the window and optimally replacing their probes using a minimum-weight perfect matching algorithm. The term independent refers to probes that can be replaced without affecting the border length of the other selected probes.

The experimental results in Kahng *et al*, 2003-1 showed that the row-epitaxial algorithm outperforms the SWM algorithm, achieving up to 9% reduction in border length when compared to the TSP-base approach of Hannenhalli *et al*, 2002.

3.1 Partitioning Algorithms

TODO: partitioning is a good idea (reduce problem size), explain Centroid-based quadrisection.

4 QUADRATIC ASSIGNMENT FORMULATION

TODO: QAP problem, probe placement as a QAP formulation, GRASP.

5 RESULTS

TODO: show results on small artificial chips.

6 DISCUSSION

TODO: extrapolate to larger chips, argue that GRASP is good as a final placer and should be combined with a partitioning algorithm; it may also be used as an optimization algorithm if modified to take into account the border problem.

REFERENCES

Chase,P.J. (1976) Subsequence numbers and logarithmic concavity, *Discrete Mathematics*, **16**, 123–140.

Fodor,S., Read,J., Pirrung,M., Stryer,L., Lu,A. and Solas,D. (1991) Light-directed, spatially addressable parallel chemical synthesis, *Science*, **251**, 767–73.

Feldman,W. and Pevzner,P. (1994) Gray code masks for sequencing by hibridization, *Genomics*, **23**, 233–235.

Hannenhalli,S., Hubell,E., Lipshutz,R. and Pevzner,P. (2002) Combinatorial algorithms for design of DNA arrays, *Advances in Biochemical Engineering / Biotechnology*, **77**, 1–19.

Kahng,A.B., Mandoiu,I.I., Pevzner,P.A., Reda,S. and Zelikovsky,A.Z. (2002) Border length minimization in DNA array design, *Proceedings of the Second Workshop on Algorithms in Bioinformatics*.

Kahng,A.B., Mandoiu,I., Pevzner,P., Reda,S. and Zelikovsky,A. (2003-1) Engineering a scalable placement heuristic for DNA probe arrays, *Proceedings of the Seventh Annual International Conference on Computational Molecular Biology*, 148–83.

Kahng, A.B., Mandoiu,I., Reda,S., Xu,X. and Zelikovsky,A. (2003-2), Evaluation of placement techniques for DNA probe array layout, *Proceedings of the IEEE/ACM International Conference on Computer-Aided Design*, 262–269.