# Improving the Layout of Oligonucleotide Microarrays: Pivot Partitioning

Sérgio A. de Carvalho Jr.[1,2,3] Sven Rahmann[1,2]

[1] Algorithms and Statistics for Systems Biology, Genome Informatics,
Technische Fakultät, Universität Bielefeld, Germany

[2] International NRW Graduate School in Bioinformatics and Genome Research

[3] Graduiertenkolleg Bioinformatik

6th Workshop on Algorithms in Bioinformatics
Zürich, Sep 2006

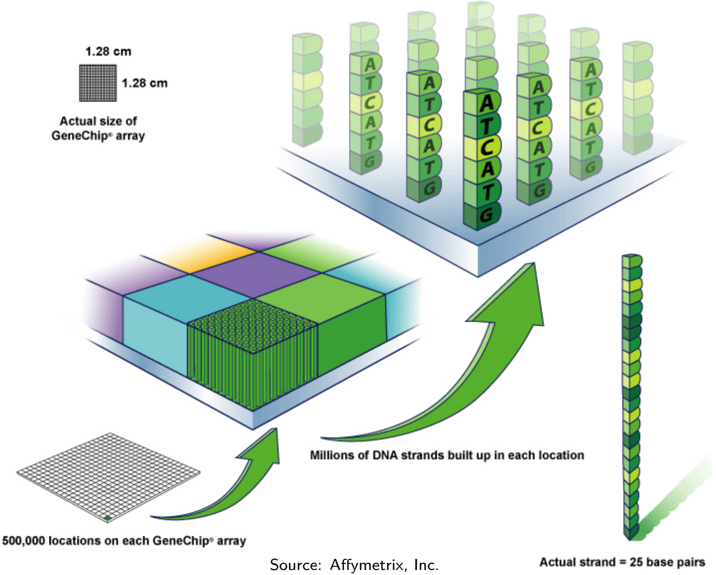
# Outline

1. Introduction: Microarray Layout
2. Conflict Index Evaluation Model
3. Pivot Partitioning Algorithm

# Outline

1. Introduction: Microarray Layout
2. Conflict Index Evaluation Model
3. Pivot Partitioning Algorithm

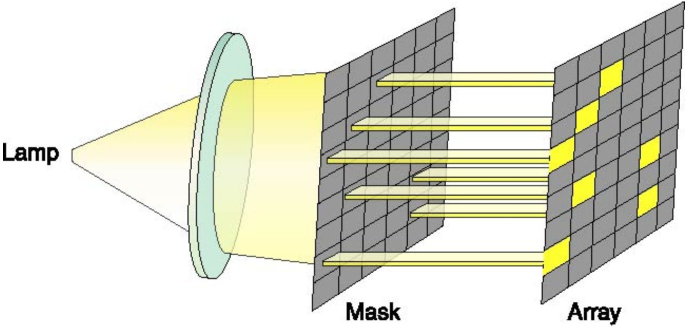
# High-Density Oligonucleotide Microarrays


1.28 cm
1.28 cm
Actual size of GeneChip® array
A T C A T G
Millions of DNA strands built up in each location
500,000 locations on each GeneChip® array
Actual strand = 25 base pairs

Source: Affymetrix, Inc.

# Probe Synthesis: Photolitographic Masks

Source: Affymetrix, Inc.



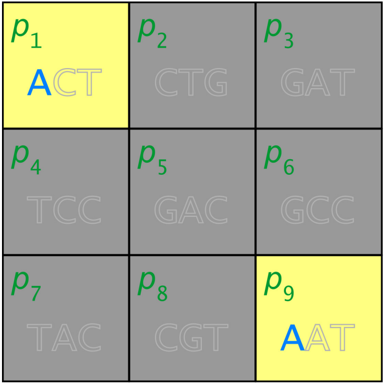
- Probes are synthesized on the chip in a series of steps
- Each step appends a particular nucleotide to selected regions
- Selection occurs by exposure to light

# Deposition Sequence and Probe Embeddings

| | | |
|---|---|---|
| $p_1$<br>ACT | $p_2$<br>CTG | $p_3$<br>GAT |
| $p_4$<br>TCC | $p_5$<br>GAC | $p_6$<br>GCC |
| $p_7$<br>TAC | $p_8$<br>CGT | $p_9$<br>AAT |

$\mathcal{S}$ = ACGTACGTACGT
$\varepsilon_1$ = ------------
$\varepsilon_2$ = ------------
$\varepsilon_3$ = ------------
$\varepsilon_4$ = ------------
$\varepsilon_5$ = ------------
$\varepsilon_6$ = ------------
$\varepsilon_7$ = ------------
$\varepsilon_8$ = ------------
$\varepsilon_9$ = ------------

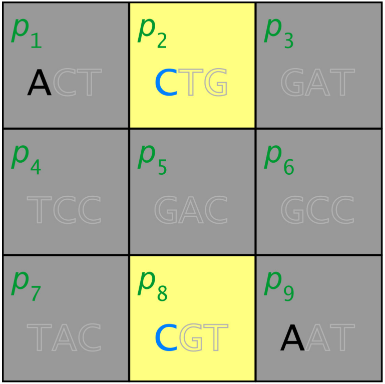
# Deposition Sequence and Probe Embeddings



| | | |
|---|---|---|
| $p_1$ ACT | $p_2$ CTG | $p_3$ GAT |
| $p_4$ TCC | $p_5$ GAC | $p_6$ GCC |
| $p_7$ TAC | $p_8$ CGT | $p_9$ AAT |

$$
\begin{aligned}
\mathcal{S} &= \text{ACGTACGTACGT}\\
\varepsilon_1 &= \text{A-----------}\\
\varepsilon_2 &= \text{------------}\\
\varepsilon_3 &= \text{------------}\\
\varepsilon_4 &= \text{------------}\\
\varepsilon_5 &= \text{------------}\\
\varepsilon_6 &= \text{------------}\\
\varepsilon_7 &= \text{------------}\\
\varepsilon_8 &= \text{------------}\\
\varepsilon_9 &= \text{A-----------}
\end{aligned}
$$

# Deposition Sequence and Probe Embeddings



| | | |
|---|---|---|
| $p_1$ ACT | $p_2$ CTG | $p_3$ GAT |
| $p_4$ TCC | $p_5$ GAC | $p_6$ GCC |
| $p_7$ TAC | $p_8$ CGT | $p_9$ AAT |

$\mathcal{S} =$ ACGTACGTACGT

$\varepsilon_1 =$ A-----------

$\varepsilon_2 =$ -C----------

$\varepsilon_3 =$ ------------

$\varepsilon_4 =$ ------------

$\varepsilon_5 =$ ------------

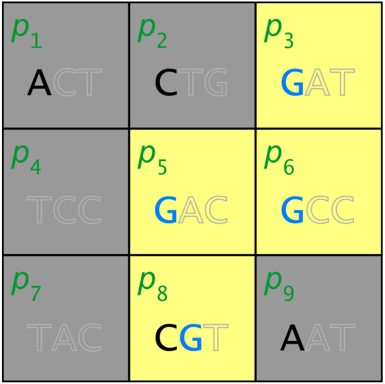
$\varepsilon_6 =$ ------------

$\varepsilon_7 =$ ------------

$\varepsilon_8 =$ -C----------

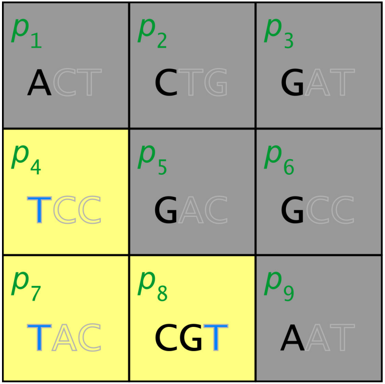
$\varepsilon_9 =$ A-----------

# Deposition Sequence and Probe Embeddings



$$\begin{aligned}
\mathcal{S} &= \text{ACGTACGTACGT}\\
\varepsilon_1 &= \text{A-----------}\\
\varepsilon_2 &= \text{-C----------}\\
\varepsilon_3 &= \text{--G---------}\\
\varepsilon_4 &= \text{------------}\\
\varepsilon_5 &= \text{--G---------}\\
\varepsilon_6 &= \text{--G---------}\\
\varepsilon_7 &= \text{------------}\\
\varepsilon_8 &= \text{-CG---------}\\
\varepsilon_9 &= \text{A-----------}
\end{aligned}$$

# Deposition Sequence and Probe Embeddings



| | | |
|---|---|---|
| $p_1$ ACT | $p_2$ CTG | $p_3$ GAT |
| $p_4$ TCC | $p_5$ GAC | $p_6$ GCC |
| $p_7$ TAC | $p_8$ CGT | $p_9$ AAT |

$$\mathcal{S} = \texttt{ACGTACGTACGT}$$
$$\varepsilon_1 = \texttt{A-----------}$$
$$\varepsilon_2 = \texttt{-C----------}$$
$$\varepsilon_3 = \texttt{--G---------}$$
$$\varepsilon_4 = \texttt{---T--------}$$
$$\varepsilon_5 = \texttt{--G---------}$$
$$\varepsilon_6 = \texttt{--G---------}$$
$$\varepsilon_7 = \texttt{---T--------}$$
$$\varepsilon_8 = \texttt{-CGT--------}$$
$$\varepsilon_9 = \texttt{A-----------}$$

# Deposition Sequence and Probe Embeddings

| | | |
|---|---|---|
| $p_1$<br>ACT | $p_2$<br>CTG | $p_3$<br>GAT |
| $p_4$<br>TCC | $p_5$<br>GAC | $p_6$<br>GCC |
| $p_7$<br>TAC | $p_8$<br>CGT | $p_9$<br>AAT |

$\mathcal{S}$ = ACGTACGTACGT

$\varepsilon_1$ = A----C-T----

$\varepsilon_2$ = -C-----T--G-

$\varepsilon_3$ = --G-A--T----

$\varepsilon_4$ = ---T-C---C--

$\varepsilon_5$ = --G-A----C--

$\varepsilon_6$ = --G--C---C--

$\varepsilon_7$ = ---TAC------

$\varepsilon_8$ = -CGT--------

$\varepsilon_9$ = A---A------T

# Deposition Sequence and Probe Embeddings

| | | |
|---|---|---|
| $p_1$<br>ACT | $p_2$<br>CTG | $p_3$<br>GAT |
| $p_4$<br>TCC | $p_5$<br>GAC | $p_6$<br>GCC |
| $p_7$<br>TAC | $p_8$<br>CGT | $p_9$<br>AAT |

$\mathcal{S}$ = ACGTACGTACGT

$\varepsilon_1$ = A----C-T----

$\varepsilon_2$ = -C-----T--G-

$\varepsilon_3$ = --G-A--T----

$\varepsilon_4$ = ---T-C---C--

$\varepsilon_5$ = --G-A----C--

$\varepsilon_6$ = --G--C---C--

$\varepsilon_7$ = ---TAC------

$\varepsilon_8$ = -CGT--------

$\varepsilon_9$ = A---A------T

$\varepsilon'_9$ = A---A--T----

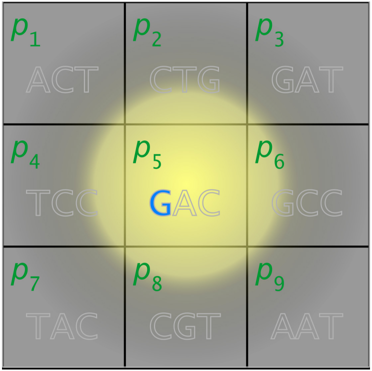
Left-most embedding!

# Deposition Sequence and Probe Embeddings

| | | |
|---|---|---|
| $p_1$<br>ACT | $p_2$<br>CTG | $p_3$<br>GAT |
| $p_4$<br>TCC | $p_5$<br>GAC | $p_6$<br>GCC |
| $p_7$<br>TAC | $p_8$<br>CGT | $p_9$<br>AAT |

$$\begin{aligned}
\mathcal{S} &= \mathrm{ACGTACGTACGT} \\
\varepsilon_1 &= 100001010000 \\
\varepsilon_2 &= 010000010010 \\
\varepsilon_3 &= 001010010000 \\
\varepsilon_4 &= 000101000100 \\
\varepsilon_5 &= 001010000100 \\
\varepsilon_6 &= 001001000100 \\
\varepsilon_7 &= 000111000000 \\
\varepsilon_8 &= 011100000000 \\
\varepsilon_9 &= 100010000001 \\
\varepsilon'_9 &= 100010010000
\end{aligned}$$

# Problem: Unintended Illumination



- Untargeted spots can be accidentally activated
    - Diffraction of light
    - Internal reflection
- Production of defective probes
- More likely near the borders between masked and unmasked spots: border conflict

**Border Length Minimization Problem (Hannenhalli et al., 2002)**

- Find arrangement (and embeddings) with minimum number of border conflicts

# Outline

1. Introduction: Microarray Layout
2. Conflict Index Evaluation Model
3. Pivot Partitioning Algorithm

# Conflict Index: Motivation

- Border Length measures the quality of a particular mask
  - We are more interested in a per-probe measure
- Practical considerations need to be taken into account:

  a) Stray light might activate probes that are as far as three cells away from the targeted spot

  b) Imperfections produced in the middle of a probe are more harmful than in its extremities

## Conflict Index of a probe $p$

$$\mathcal{C}(p) := \sum_{t=1}^{T} \left( \omega(p,t) \sum_{p'} \delta(p,p',t) \right),$$

where $\delta(p,p',t)$ are distance-dependent weights (a) and $\omega(p,t)$ are position-dependent weights (b) defined as follows.

# Conflict Index: Definition

## Conflict Index of a probe $p$

$$\mathcal{C}(p) := \sum_{t=1}^{T} \left( \omega(p,t) \sum_{p'} \delta(p,p',t) \right)$$

| | | | | | | |
|---|---|---|---|---|---|---|
| 0.06 | 0.08 | 0.10 | 0.11 | 0.10 | 0.08 | 0.06 |
| 0.08 | 0.13 | 0.20 | 0.25 | 0.20 | 0.13 | 0.08 |
| 0.10 | 0.20 | 0.50 | 1.00 | 0.50 | 0.20 | 0.10 |
| 0.11 | 0.25 | 1.00 | $p$ | 1.00 | 0.25 | 0.11 |
| 0.10 | 0.20 | 0.50 | 1.00 | 0.50 | 0.20 | 0.10 |
| 0.08 | 0.13 | 0.20 | 0.25 | 0.20 | 0.13 | 0.08 |
| 0.06 | 0.08 | 0.10 | 0.11 | 0.10 | 0.08 | 0.06 |

## a) Distance-dependent weights $\delta(p, p', t)$

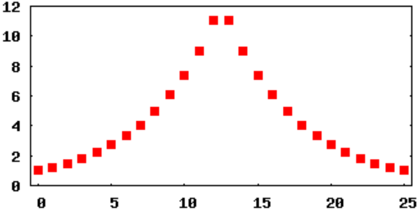
$$\delta(p,p',t) := \begin{cases} (d(p,p'))^{-2} & \text{if } p' \text{ is unmasked at step } t, \\ 0 & \text{otherwise,} \end{cases}$$

where $d(p,p')$ is the Euclidean distance between the spots of $p$ and $p'$.

# Conflict Index: Definition

## Conflict Index of a probe $p$

$$\mathcal{C}(p) := \sum_{t=1}^{T}\Big(\omega(p,t)\sum_{p'}\delta(p,p',t)\Big)$$



## b) Position-dependent weights $\omega(p,t)$

$$\omega(p,t) := \begin{cases} c \cdot \exp\left(\theta \cdot \lambda(p,t)\right) & \text{if } p \text{ is masked at step } t, \\ 0 & \text{otherwise,} \end{cases}$$

where $c > 0$ and $\theta > 0$ are constants,

$$\lambda(p,t) := 1 + \min(b_{p,t}, \ell_p - b_{p,t}),$$

$b_{p,t}$ denotes the number of nucleotides synthesized up to and including step $t$, and $\ell_p$ is the length of probe $p$.

# Outline

1. Introduction: Microarray Layout

2. Conflict Index Evaluation Model

3. Pivot Partitioning Algorithm

# Previous Work: Place and Re-embed

- The microarray layout problem has been traditionally approached in two phases:
  1) Placement of probes given a fixed embedding
  2) Re-embedding of probes given a fixed placement

## Placement: Row-epitaxial (Kahng *et al.*, 2003)

- Essentially greedy
- Spots are filled in a pre-defined order
  - Select probe from a list $Q$ such that conflicts with filled spots are minimized
- Restrict the maximum size of $Q$

# Previous Work: Place and Re-embed

## Re-embedding: several algorithms

- All based on the Optimum Single Probe Embedding (OSPE)
- OSPE re-embed a probe optimally in regards to its neighbors
- Difference is in the order in which re-embeddings take place

$\mathcal{S}$ = ACGTACGTACGT

$\varepsilon_1$ = A----C-T----

$\varepsilon_2$ = -C-----T--G-

$\varepsilon_3$ = --G-A--T----

$\varepsilon_4$ = ---T-C---C--

$\varepsilon_5$ = ????????????

$\varepsilon_6$ = --G--C---C--

$\varepsilon_7$ = ---TAC------

$\varepsilon_8$ = -CGT--------

$\varepsilon_9$ = A---A------T

## Optimum Single Probe Embedding (OSPE)

- Dynamic Programming
- Originally developed for border length minimization
- Now extended for conflict index minimization

# Previous Work: Partitioning

More recently, a partitioning algorithm was proposed

- Divide the problem into smaller sub-problems
- Each sub-problem is treated as a separate placement
- Reduce run-time; may improve placement

Partitioning: Centroid-based Quadrisection (Kahng *et al.*, 2003)

- To do...

# Summary

- Conflict Index: new model for evaluating microarray layouts
- Pivot Partitioning: new partitioning algorithm
    - Faster and better selection of pivots
    - Improved assignment of probes to regions
    - First to combine placement and re-embedding

# Thanks!

GI Universität Bielefeld
Technische Fakultät
AG Genominformatik

International Graduate School
Bioinformatics & Genome Research

- Prof. Dr. Jens Stoye
- Prof. Dr. Robert Giegerich
- AG Genominformatik
- Graduiertenkolleg Bioinformatik
- Graduate School in Bioinformatics and Genome Research
- ...and thank you for your attention!

More info on

http://gi.cebitec.uni-bielefeld.de/assb/chiplayout