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

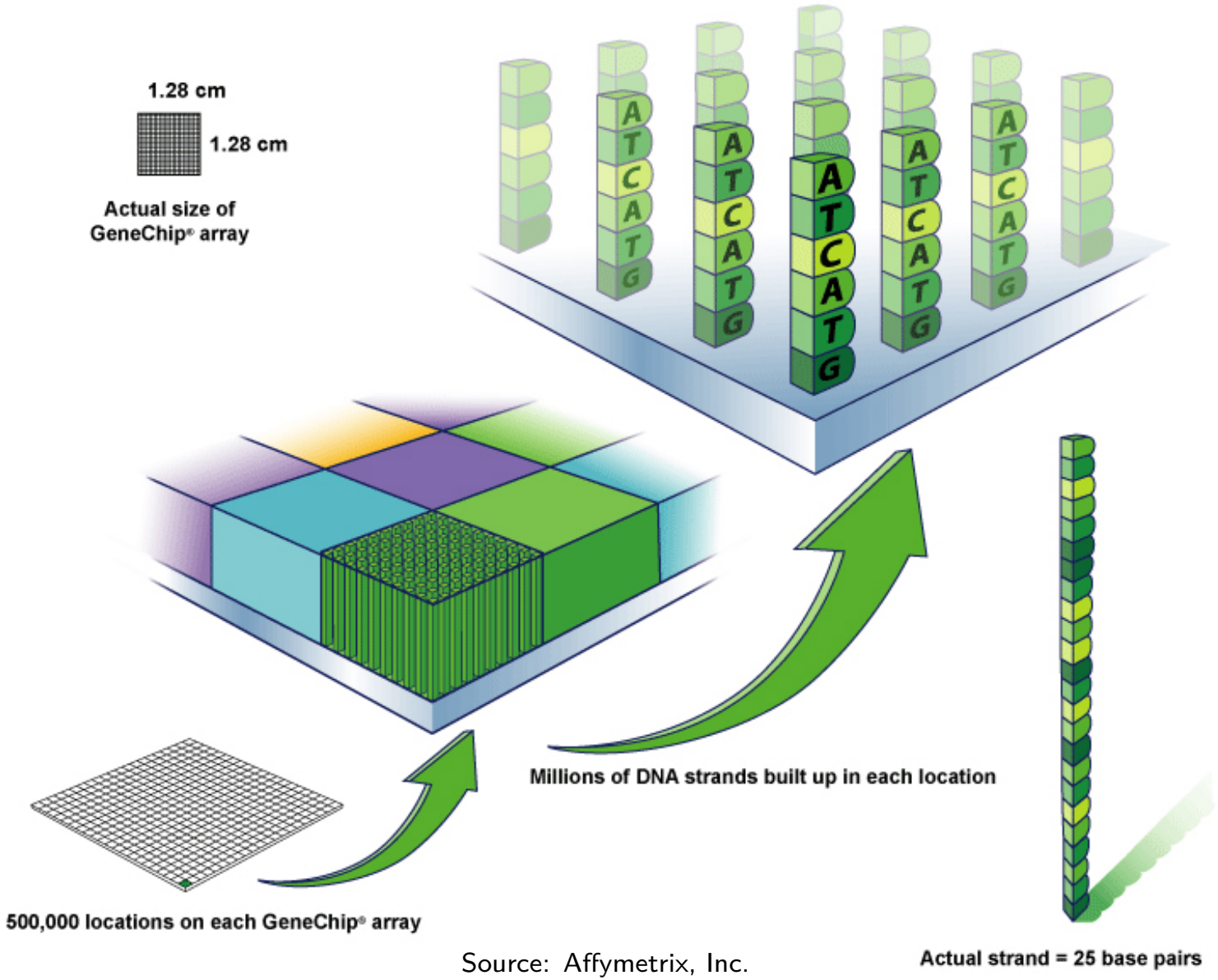
[3]Graduiertenkolleg Bioinformatik

Workshop on Algorithms in Bioinformatics, 2006

Outline

Outline

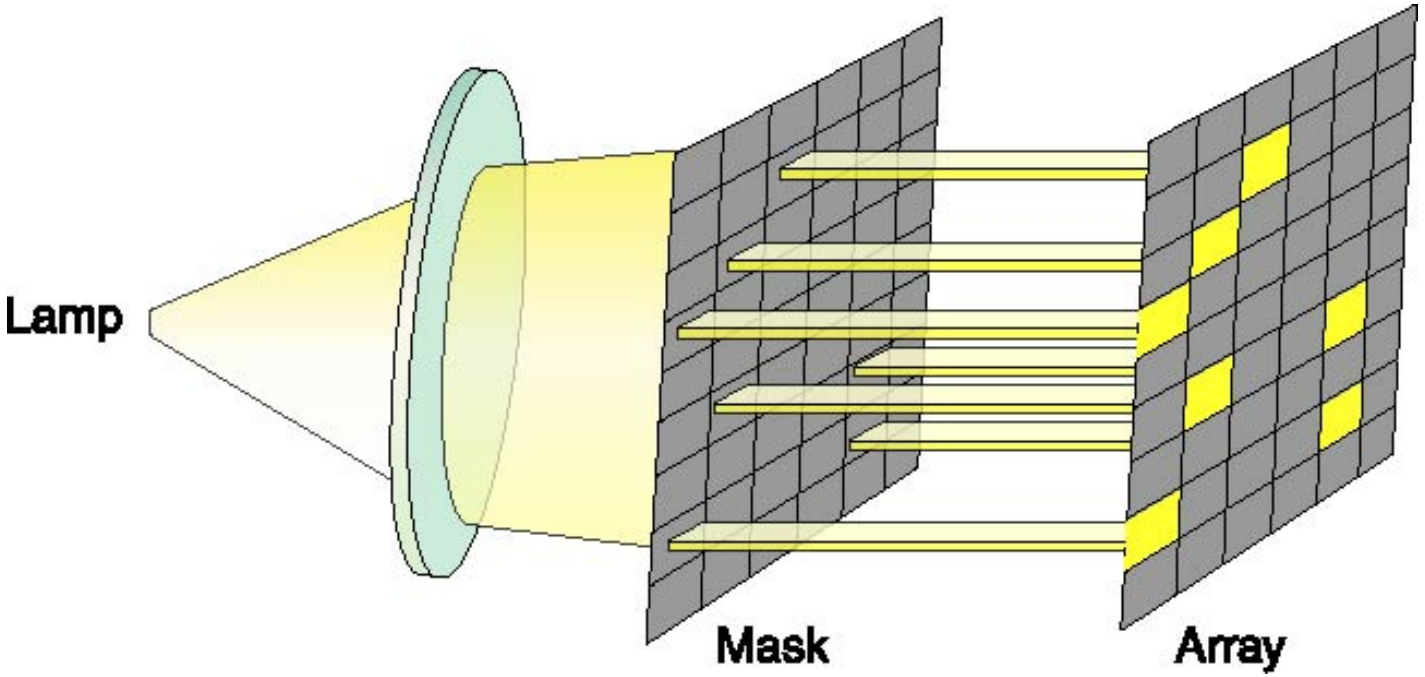
High-Density Oligonucleotide Microarrays



Source: Affymetrix, Inc.

Probe Synthesis with Photolitographic Masks

Source: Affymetrix, Inc.



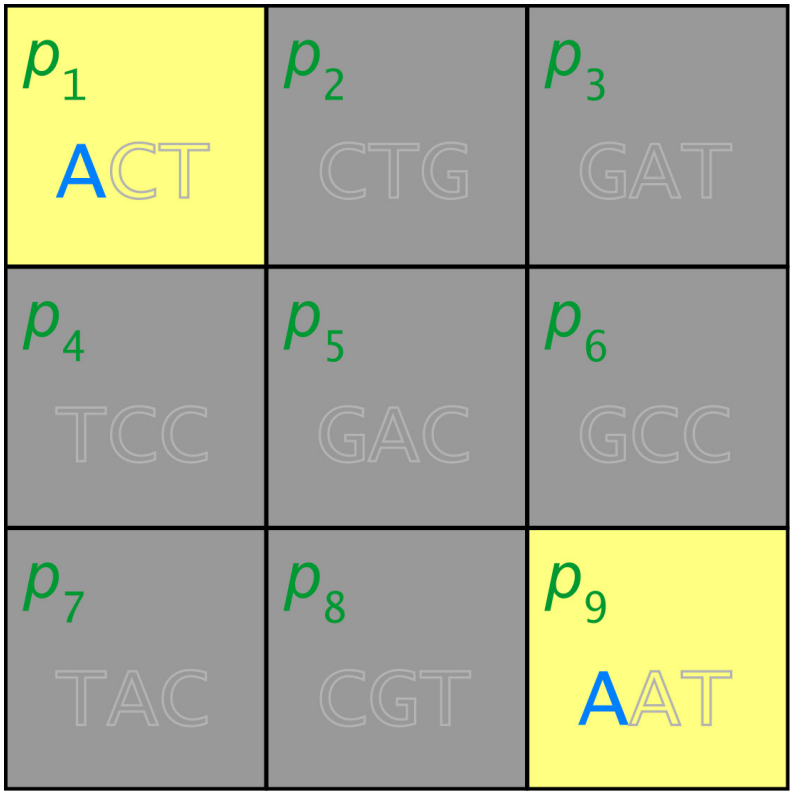
- Probes are synthesized on the chip in a series of steps
- Each step appends a particular nucleotide to selected regions
- Selection occurs by exposure to light directed by a mask

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

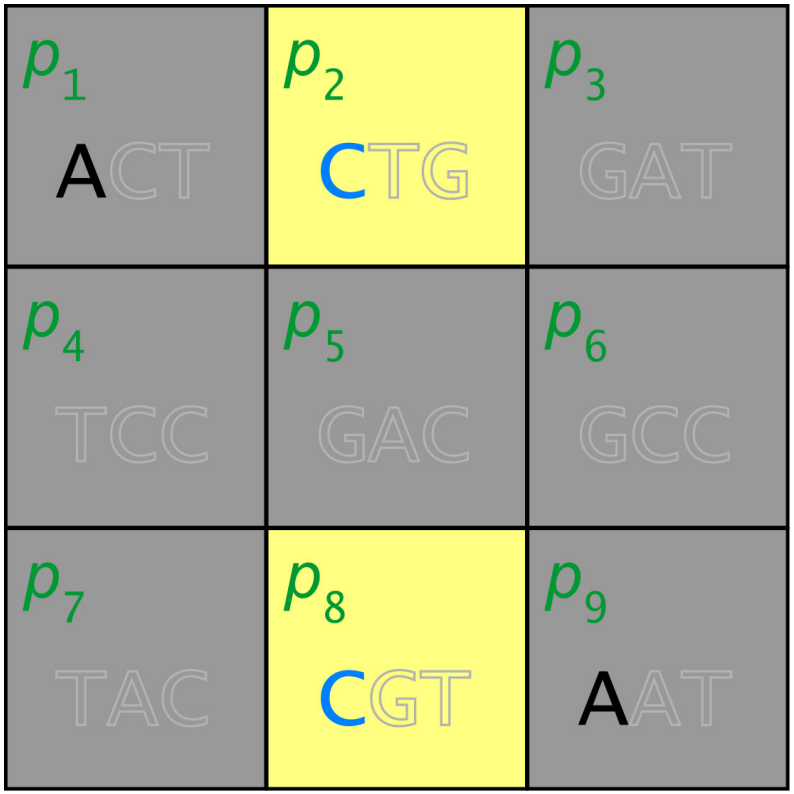
$$
\begin{aligned}
S &= \text{ACGTACGTACGT} \\
\varepsilon_1 &= \text{------------} \\
\varepsilon_2 &= \text{------------} \\
\varepsilon_3 &= \text{------------} \\
\varepsilon_4 &= \text{------------} \\
\varepsilon_5 &= \text{------------} \\
\varepsilon_6 &= \text{------------} \\
\varepsilon_7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

Deposition Sequence and Probe Embeddings



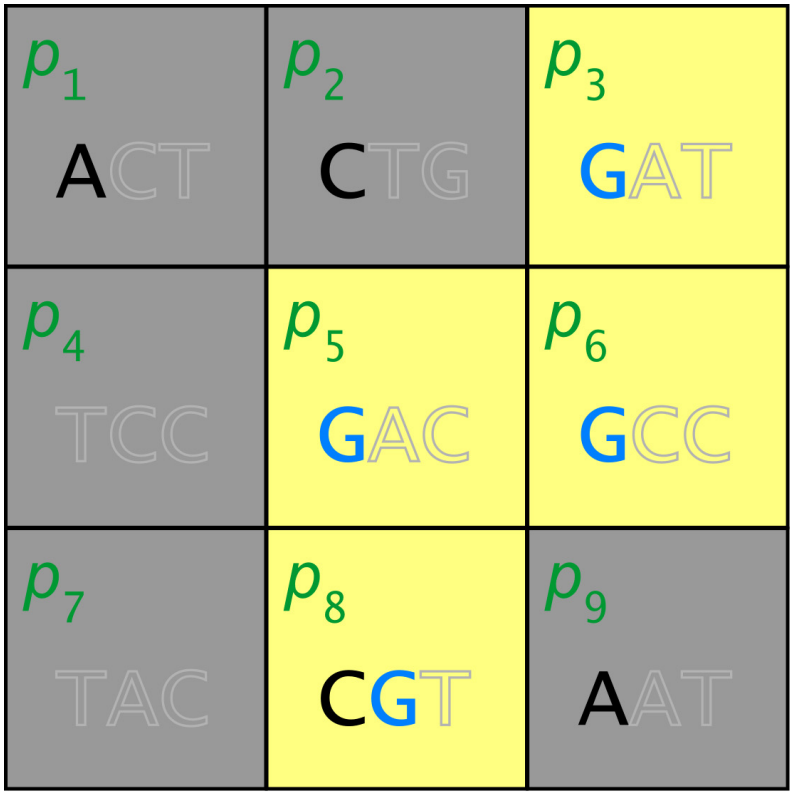
$$
\begin{aligned}
S &= \mathtt{ACGTACGTACGT} \\
\varepsilon_1 &= \mathtt{A-----------} \\
\varepsilon_2 &= \mathtt{------------} \\
\varepsilon_3 &= \mathtt{------------} \\
\varepsilon_4 &= \mathtt{------------} \\
\varepsilon_5 &= \mathtt{------------} \\
\varepsilon_6 &= \mathtt{------------} \\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

Deposition Sequence and Probe Embeddings



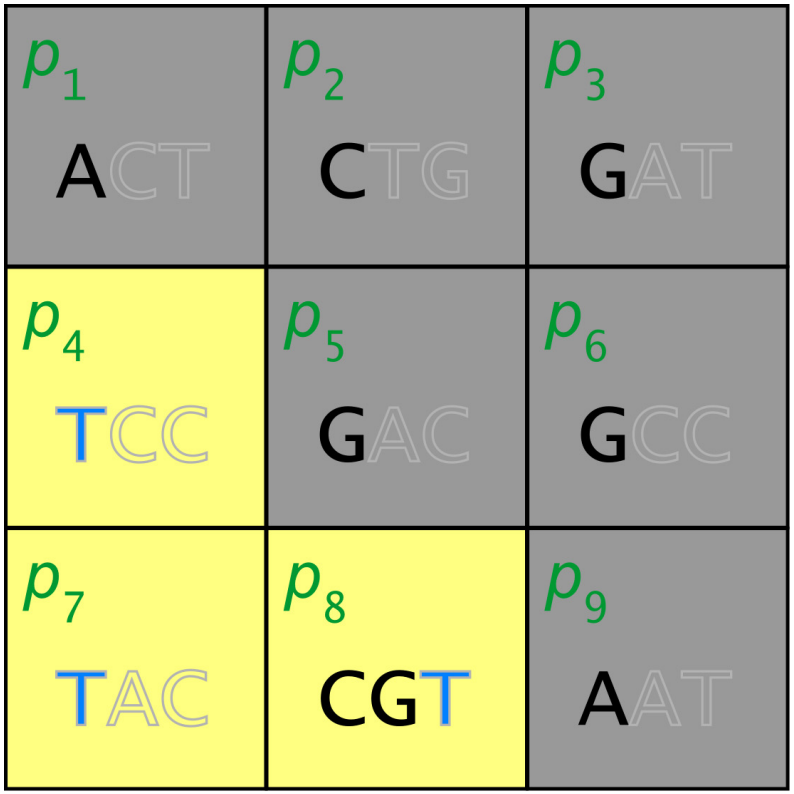
$$
\begin{aligned}
S &= \text{ACGTACGTACGT} \\
\varepsilon_1 &= \text{A----------} \\
\varepsilon_2 &= \text{-C---------} \\
\varepsilon_3 &= \text{----------} \\
\varepsilon_4 &= \text{----------} \\
\varepsilon_5 &= \text{----------} \\
\varepsilon_6 &= \text{----------} \\
\varepsilon_7 &= \text{----------}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

Deposition Sequence and Probe Embeddings



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

Deposition Sequence and Probe Embeddings



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

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

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

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

S = ACGTACGTACGT

ε_1 = A----C-T----

ε_2 = -C-----T--G-

ε_3 = --G-A--T----

ε_4 = ---T-C---C--

ε_5 = --G-A----C--

ε_6 = --G--C---C--

ε_7 = ---TAC------

ε_8 = -CGT--------

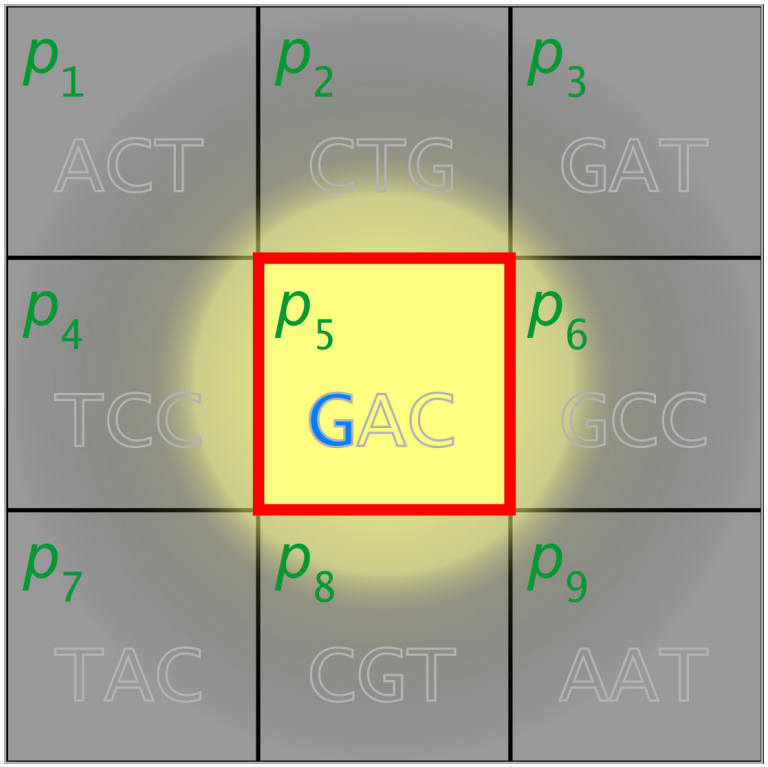
ε_9 = A---A------T

ε'_9 = A-------A--T

Right-most: ε''_9 = ----A---A--T

Left-most: ε'''_9 = A---A--T----

Unintended Illumination Problem



- 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

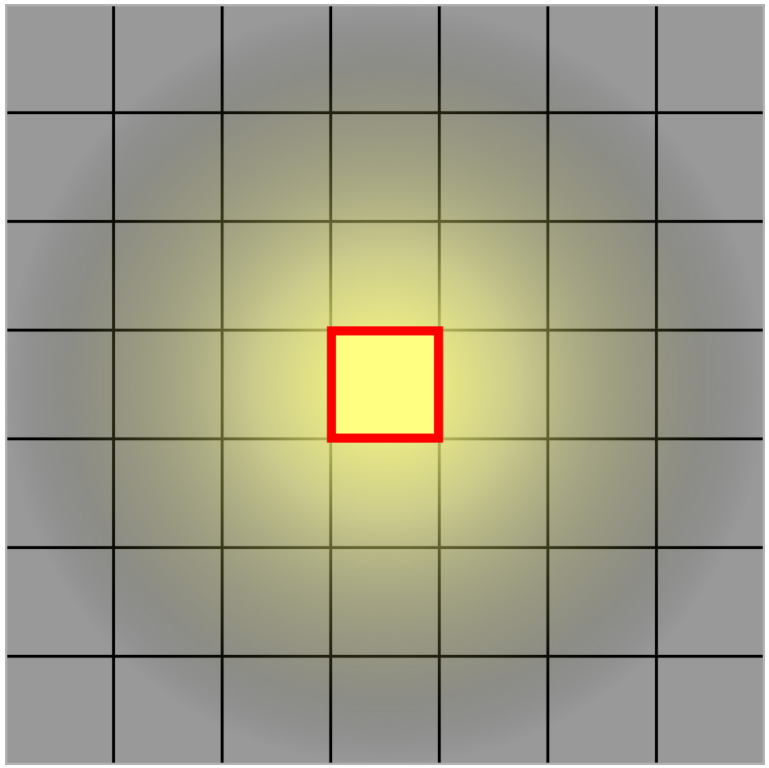
Border Length Minimization Problem (Hannenhalli et al., 2002)

Find arrangement of the probes and embeddings with minimum number of border conflicts over all masks

Outline

Motivation

- Border Length measures the quality of a particular mask
  - We are more interested in a per-probe measure
- Practical considerations:
  a) Stray light might damage probes as far as three cells away from the targeted spot
  b) Imperfections in the middle of a probe are more harmful than in its extremities



ATGACTACCATGCAGTACAACATAC

Definition

Conflict Index of a probe p

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

Distance-dependent weights

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

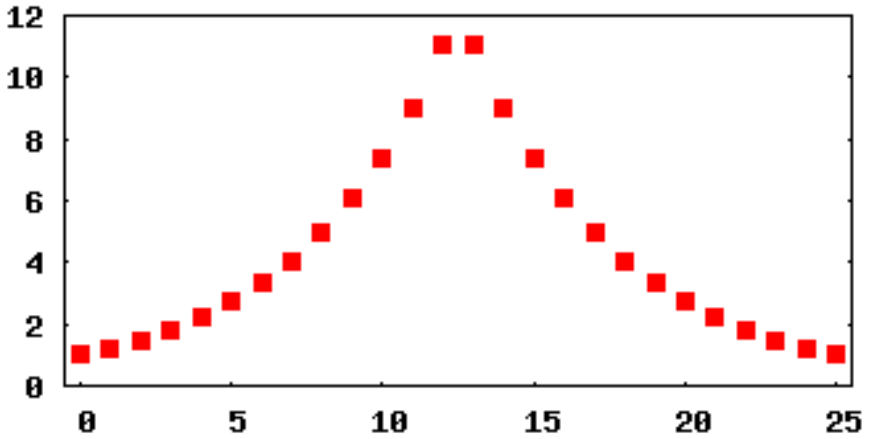
where $d(p,p')$ is the Euclidean distance between the spots of p and p'.

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

Definition

Conflict Index of a probe p

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



Position-dependent weights

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

where

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

$b_{p,t}$ denotes the number of nucleotides synthesized up to and including step t, ℓ_p is the length of probe p, $c > 0$ and $\theta > 0$ are constants.

New Problem

Conflict Index Minimization Problem

Find placement of the probes and embeddings such that

$$\sum_{p} \mathcal{C}(p) \rightarrow \min$$

Outline

Previous Work: Place and Re-embed

The problem has been traditionally approached in two phases:

1) Placement of probes given a fixed embedding
2) Re-embedding of probes once a placement is fixed

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

- 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 (e.g. $Q = 20\,000$)

Re-embedding: Sequential (Kahng *et al.*, 2003)

- Based on the Optimum Single Probe Embedding (OSPE)
  - Re-embed a probe optimally in regards to its neighbors
- Spots are re-embedded in a pre-defined order

Optimum Single Probe Embedding (OSPE)

- Designed for border length minimization (Kahng *et al.*,2002)
  - We extended it for conflict index minimization

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

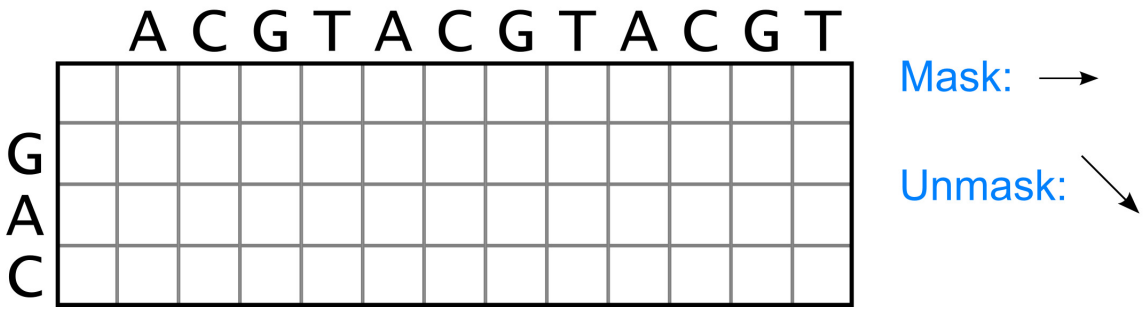
$$
\begin{aligned}
S &= \text{ACGTACGTACGT} \\
\varepsilon_1 &= \text{A----C-T----} \\
\varepsilon_2 &= \text{-C-----T--G-} \\
\varepsilon_3 &= \text{--G-A--T----} \\
\varepsilon_4 &= \text{---T-C---C--} \\
\varepsilon_5 &= ?????????? \\
\varepsilon_6 &= \text{--G--C---C--} \\
\varepsilon_7 &= \text{---TAC------}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

Optimum Single Probe Embedding (OSPE)

- Designed for border length minimization (Kahng *et al.*,2002)
  - We extended it for conflict index minimization

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

S = ACGTACGTACGT
ε_1 = A----C-T----
ε_2 = -C-----T--G-
ε_3 = --G-A--T----
ε_4 = ---T-C---C--
ε_5 = ??????????
ε_6 = --G--C---C--
ε_7 = ---TAC------
ε_8 = -CGT--------
ε_9 = A---A------T

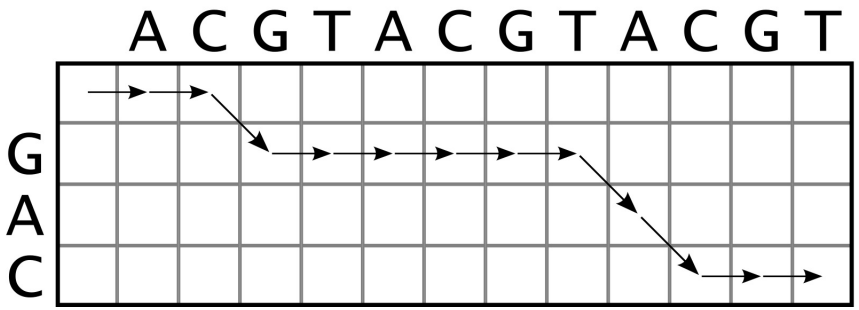


Optimum Single Probe Embedding (OSPE)

- Designed for border length minimization (Kahng *et al.*,2002)
  - We extended it for conflict index minimization

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

S = ACGTACGTACGT
ε_1 = A----C-T----
ε_2 = -C-----T--G-
ε_3 = --G-A--T----
ε_4 = ---T-C---C--
ε_5 = --G-----AC--
ε_6 = --G--C---C--
ε_7 = ---TAC------
ε_8 = -CGT--------
ε_9 = A---A------T



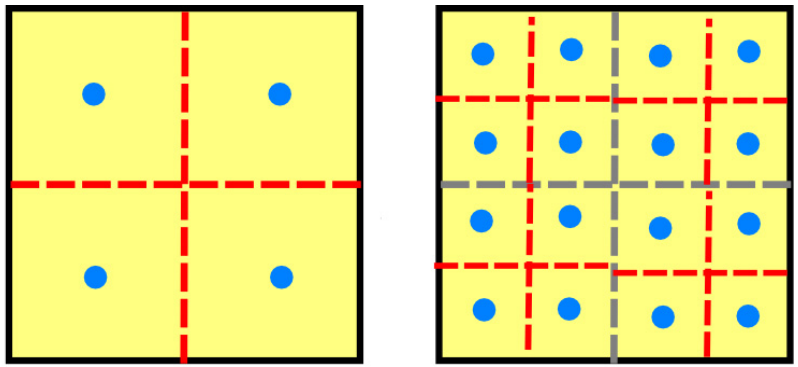
Partitioning

- The placement problem can be partitioned
  - Divide the chip into sub-regions; assign sub-sets of probes to each sub-region
  - Sub-regions are processed independently, and can be recursively partitioned
  - A placement algorithm is called on each final sub-region
- Partitioning is a compromise but can reduce run-time and help the placement algorithm

Partitioning

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

- Heuristically select four probes maximizing the Hamming distance between their embeddings: centroids
- Chip is divided into four quadrants, each gets one centroid
- Remaining probes p are assigned to the quadrant whose centroid has the minimum Hamming distance to p



What's New?

ACGTACGTACGTACGTACGTACGTACGT
--G--C--A---A--T--G--C--A-G-
AC-TA-G-ACG-A----------------

Observation 1

- Some probes have only a few possible embeddings
  - Their embeddings cover most cycles of the deposition sequence
- Others may have up to several million embeddings
  - Particular embeddings may not serve as good centroids

Pivot Partitioning

- Restrict the selection of centroids to probes with fewer embeddings: pivots
  - Faster and better selection of centroids
  - Probes with more embeddings can more "easily" adapt to the pivots

What's New?

Observation 2

- The place and re-embed approach is inefficient
  - Placement is based on embeddings that are likely to change

Pivot Partitioning

- Consider all embeddings when assigning probes to sub-regions (using OSPE)
  - Better assignment of probes
- Re-embed probes in regards to the pivots before placement (OSPE again)
  - Improve the "alignment" of the embeddings in the region

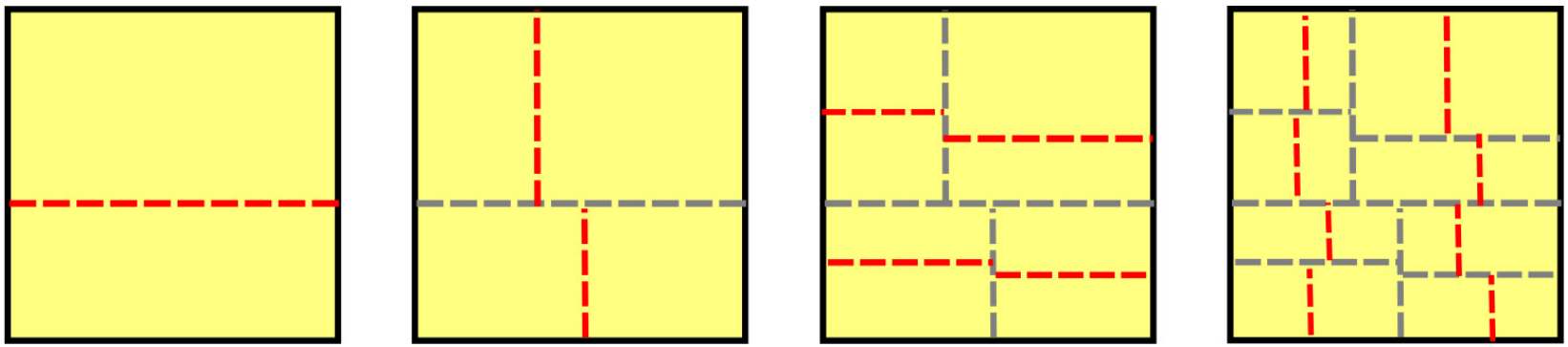
What's New?

Observation 3

- The quadrisection partitioning may force bad assignments in order to ensure squared regions

Pivot Partitioning

- Alternate horizontal and vertical partitions
- Allow sub-regions to have different sizes
  - Greater flexibility when assingning probes to sub-regions



Results on Random Chips

Using Row-epitaxial for the placement (with $Q = 20\,000$), followed by Sequential re-embedding

Border Length Minimization

Dim	$t_{max} = 0$		$t_{max} = 2$		$t_{max} = 4$		$t_{max} = 6$	
	Cost	Time	Cost	Time	Cost	Time	Cost	Time
100×100	42.77	34	39.19	13	40.72	10	42.11	11
200×200	41.63	429	37.30	155	38.53	62	40.00	85
300×300	41.38	1 174	36.12	766	37.22	264	38.53	139
500×500	41.27	3 524	34.69	3 472	35.50	1 996	36.58	713

t_{max}: maximum partitioning depth
Dim: chip dimension
Cost: normalized border length
Time: running time in seconds

Results on Random Chips

Using Row-epitaxial for the placement (with $Q = 2\,000$), followed by Sequential re-embedding

Conflict Index minimization

Dim	$t_{max} = 0$ Cost	$t_{max} = 0$ Time	$t_{max} = 2$ Cost	$t_{max} = 2$ Time	$t_{max} = 4$ Cost	$t_{max} = 4$ Time	$t_{max} = 6$ Cost	$t_{max} = 6$ Time
100×100	514.49	45	453.67	37	467.78	19	475.44	15
200×200	517.07	192	466.22	215	452.41	166	462.55	99
300×300	518.51	438	475.84	524	452.00	466	448.17	336
500×500	517.50	1 471	481.36	1 530	462.33	1 472	445.43	1 295

t_{max}: maximum partitioning depth
Dim: chip dimension
Cost: average conflict index
Time: running time in seconds

Pivot Partitioning vs. Centroid Quadrisection

Using Row-epitaxial for the placement (with $Q = 20\,000$), followed by Sequential re-embedding

Border Length Minimization

Dim	$L = 1$ CQ	$t_{max} = 2$ PP	$L = 2$ CQ	$t_{max} = 4$ PP	$L = 3$ CQ	$t_{max} = 6$ PP
100×100	393 218	0.18%	399 312	-1.89%	410 608	-2.48%
200×200	1 524 803	2.27%	1 545 825	0.48%	1 573 096	-1.34%
300×300	3 493 552	7.12%	3 413 316	2.05%	3 434 964	-0.61%
500×500	9 546 351	8.95%	9 355 231	4.67%	9 307 510	1.03%

L: maximum partitioning depth of Centroid Quadrisection (CQ)
t_{max}: maximum partitioning depth of Pivot Partitioning (PP)
Dim: chip dimension
Cost: normalized border length

- CQ's "borrowing heuristic" may explain why it performs better on small chips and deeper partitioning levels
- Running times are in the same order of magnitude (not shown)

Summary

- Conflict Index
  - New model for evaluating microarray layouts
- Pivot Partitioning
  - First algorithm to combine placement and re-embedding
  - Improved assignment of probes to regions
  - Up to 8.95% reduction of conflicts compared to existing algorithms with comparable running times
- Future work
  - Implement “borrowing heuristic” for Pivot Partitioning

Auf Wiedersehen!

To appear...

- Microarray Layout as a Quadratic Assignment Problem, German Conference on Bioinformatics (GCB 2006), LNI.
- New placement algorithm under development...

More info on

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

- Thanks to Ion Mandoiu, Xu Xu and Sherif Reda
- And thank you for your attention!