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

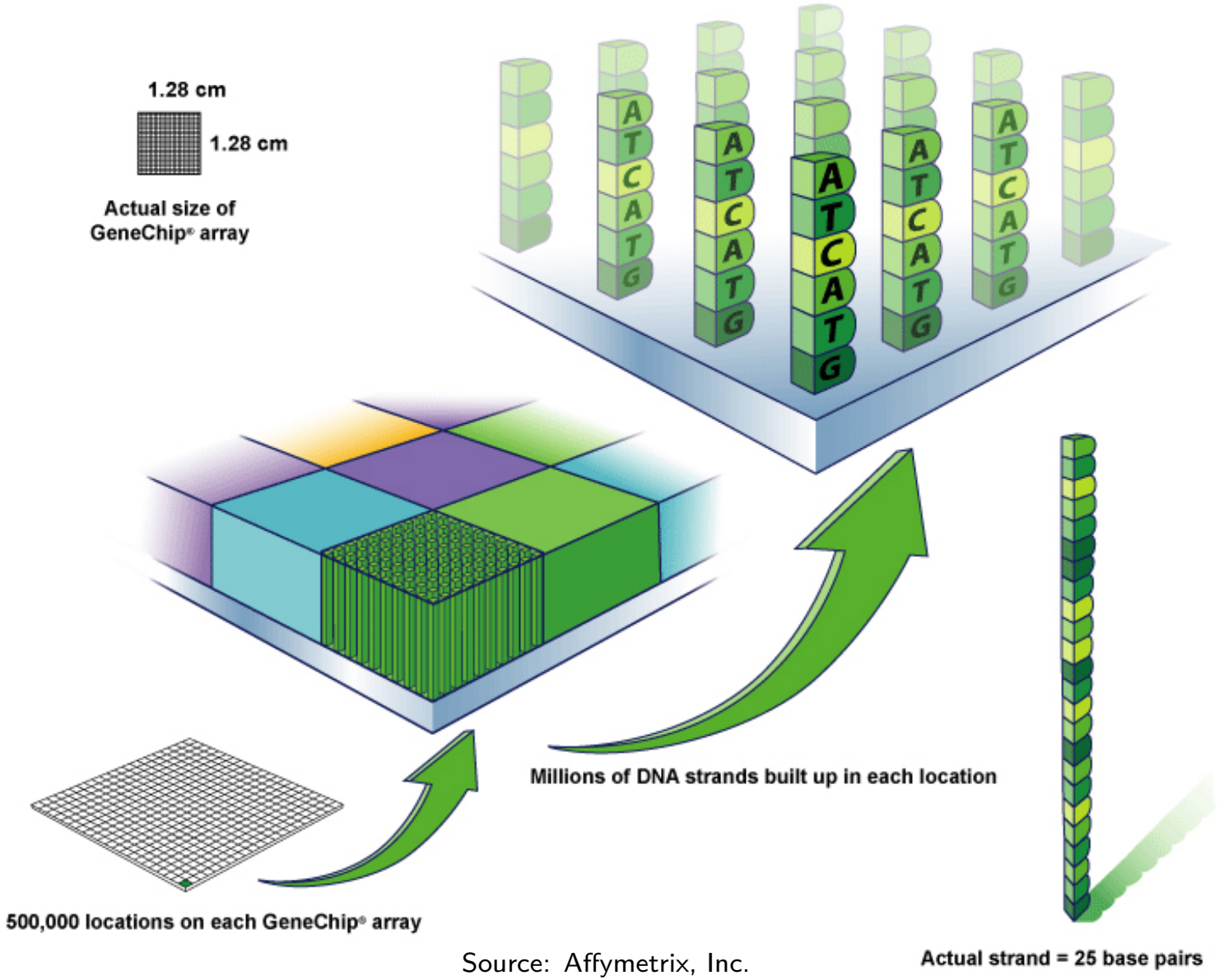
[3]Graduiertenkolleg Bioinformatik

Workshop on Algorithms in Bioinformatics, 2006

Outline

Outline

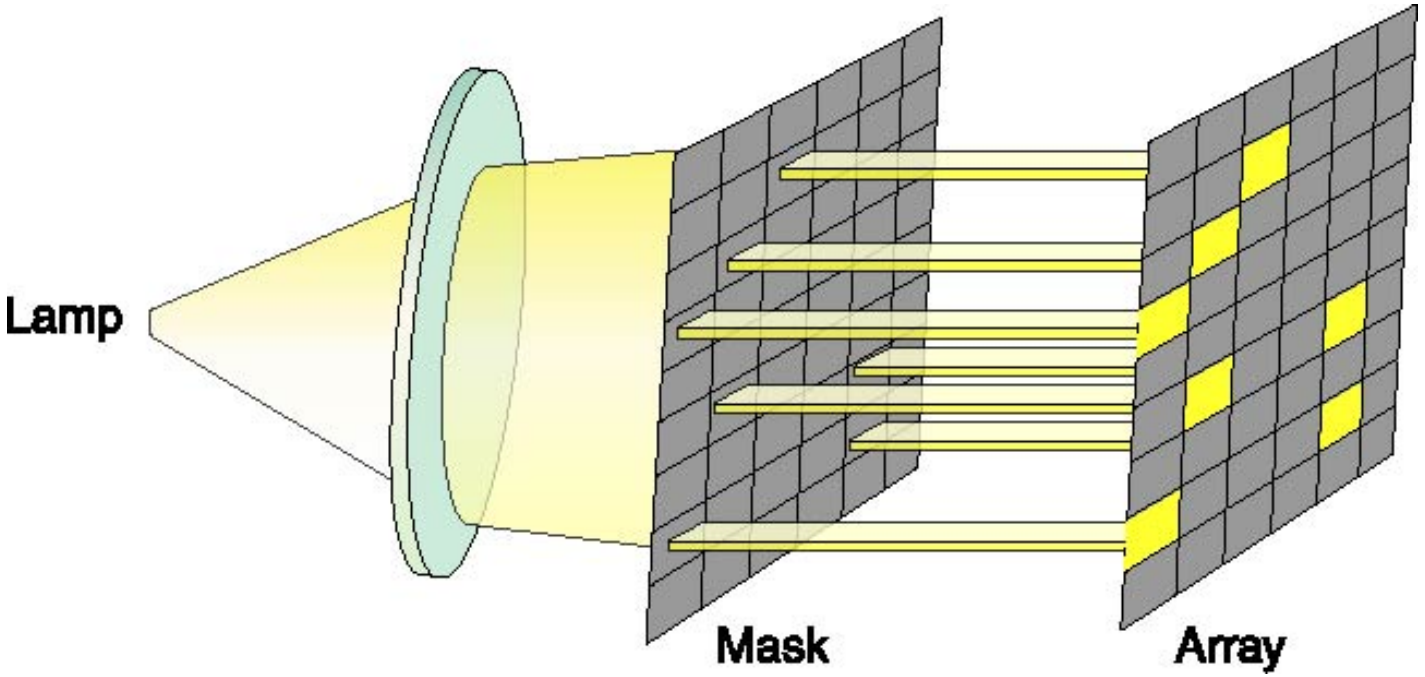
High-Density Oligonucleotide Microarrays



Source: Affymetrix, Inc.

Probe Synthesis with Photolitographic Masks

Source: Affymetrix, Inc.



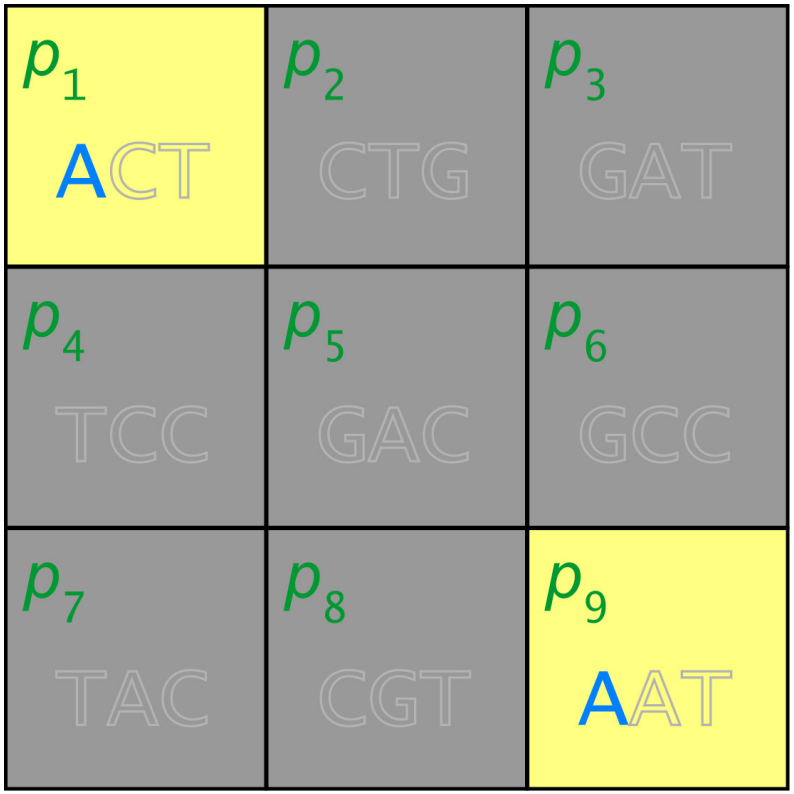
- Probes are synthesized on the chip in a series of steps
- Each step appends a particular nucleotide to selected regions
- Selection occurs by exposure to light directed by a mask

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

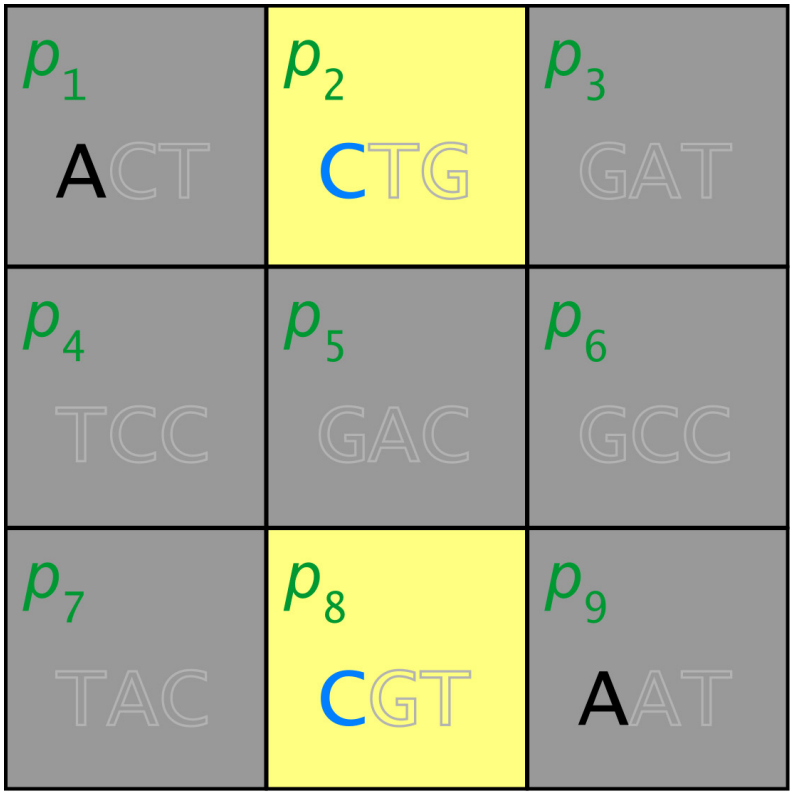
$$
\begin{aligned}
S &= \text{ACGTACGTACGT} \\
\varepsilon_1 &= \text{------------} \\
\varepsilon_2 &= \text{------------} \\
\varepsilon_3 &= \text{------------} \\
\varepsilon_4 &= \text{------------} \\
\varepsilon_5 &= \text{------------} \\
\varepsilon_6 &= \text{------------} \\
\varepsilon_7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

Deposition Sequence and Probe Embeddings



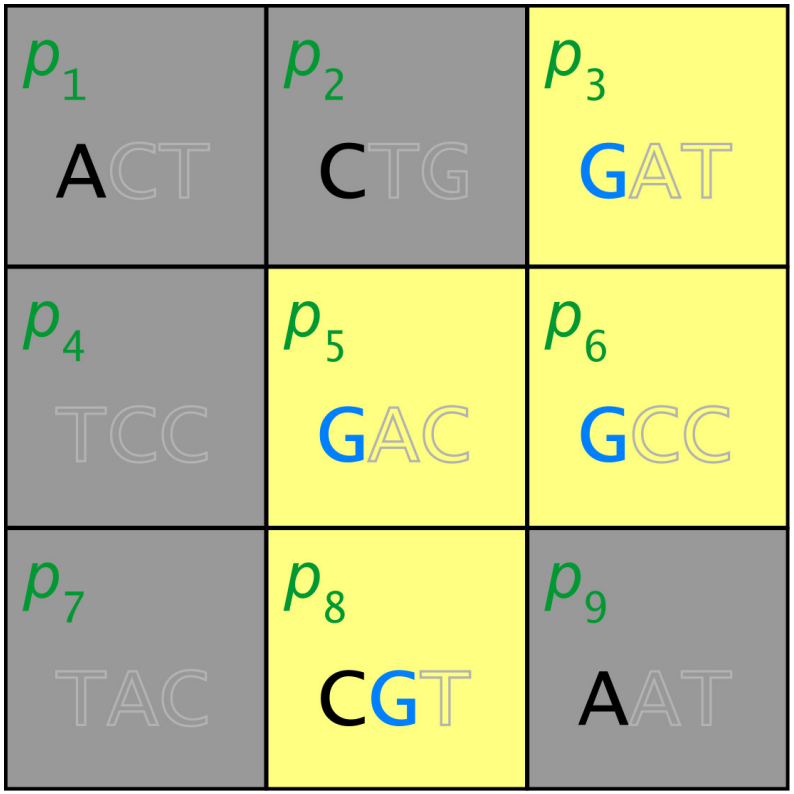
$$
\begin{aligned}
S &= \mathtt{ACGTACGTACGT} \\
\varepsilon_1 &= \mathtt{A-----------} \\
\varepsilon_2 &= \mathtt{------------} \\
\varepsilon_3 &= \mathtt{------------} \\
\varepsilon_4 &= \mathtt{------------} \\
\varepsilon_5 &= \mathtt{------------} \\
\varepsilon_6 &= \mathtt{------------} \\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

Deposition Sequence and Probe Embeddings



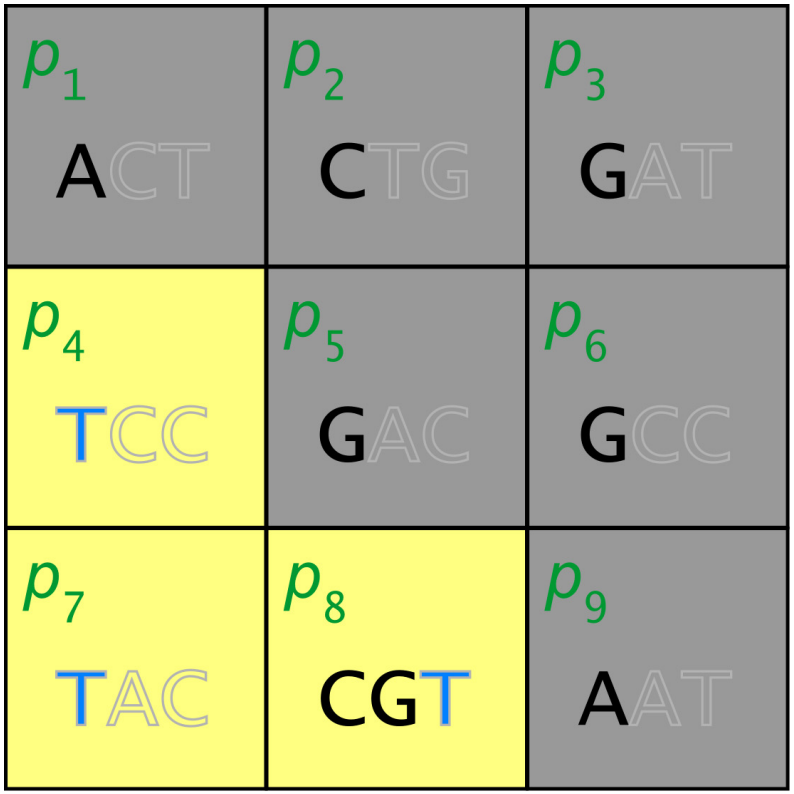
$$
\begin{aligned}
S &= \text{ACGTACGTACGT} \\
\varepsilon_1 &= \text{A----------} \\
\varepsilon_2 &= \text{-C----------} \\
\varepsilon_3 &= \text{----------} \\
\varepsilon_4 &= \text{----------} \\
\varepsilon_5 &= \text{----------} \\
\varepsilon_6 &= \text{----------} \\
\varepsilon_7 &= \text{----------}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

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

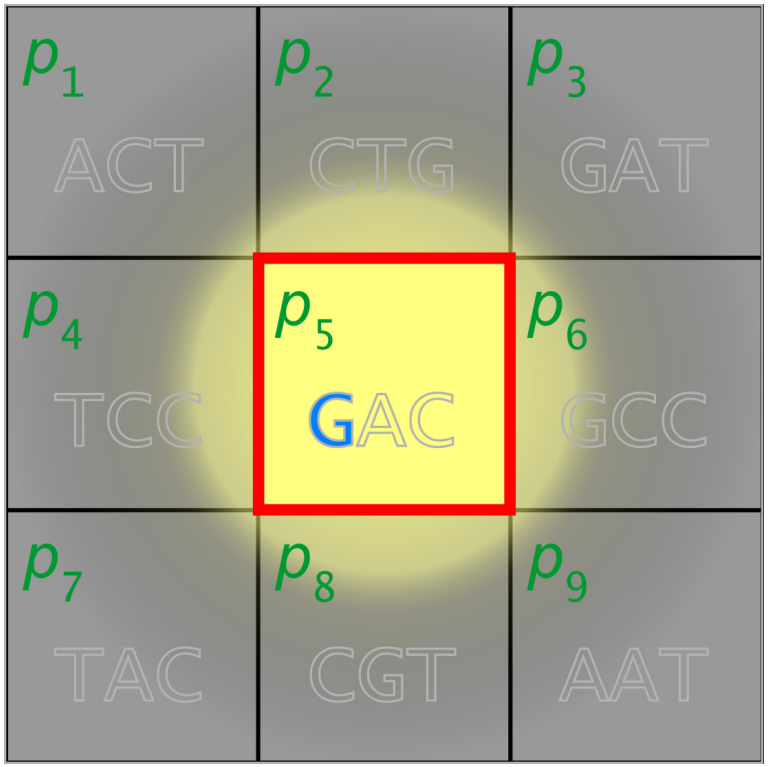
$$
\begin{aligned}
S &= \text{ACGTACGTACGT} \\
\varepsilon_1 &= \text{A----C-T----} \\
\varepsilon_2 &= \text{-C------T--G-} \\
\varepsilon_3 &= \text{--G-A--T----} \\
\varepsilon_4 &= \text{---T-C---C--} \\
\varepsilon_5 &= \text{--G-A----C--} \\
\varepsilon_6 &= \text{--G--C---C--} \\
\varepsilon_7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

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

$$
\begin{aligned}
S &= \text{ACGTACGTACGT} \\
\varepsilon_1 &= \text{A----C-T----} \\
\varepsilon_2 &= \text{-C-----T--G-} \\
\varepsilon_3 &= \text{--G-A--T----} \\
\varepsilon_4 &= \text{---T-C---C--} \\
\varepsilon_5 &= \text{--G-A----C--} \\
\varepsilon_6 &= \text{--G--C---C--} \\
\varepsilon_7 &= \text{---TAC------} \\
\varepsilon_8 &= \text{-CGT--------} \\
\varepsilon_9 &= \text{A---A------T} \\
\varepsilon'_9 &= \text{A-------A--T} \\
\text{Right-most: } \varepsilon''_9 &= \text{----A---A--T} \\
\text{Left-most: } \varepsilon'''_9 &= \text{A---A--T----}
\end{aligned}
$$

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

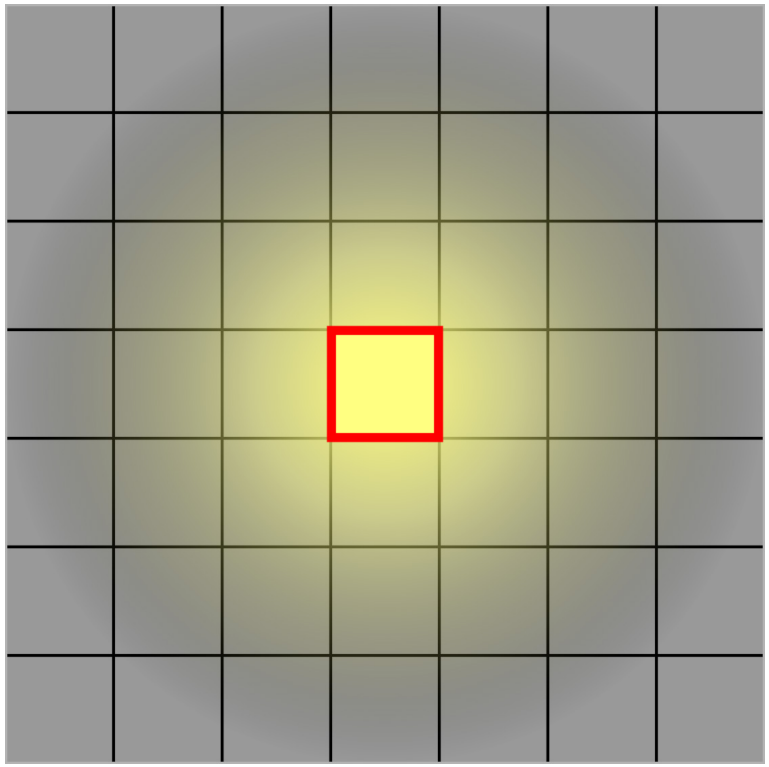
Border Length Minimization Problem (Hannenhalli et al., 2002)

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

Outline

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 damage probes that lie 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



ATGACTACCATGCAGTACAACATAC

Definition

Conflict Index of a probe p

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

Distance-dependent weights

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

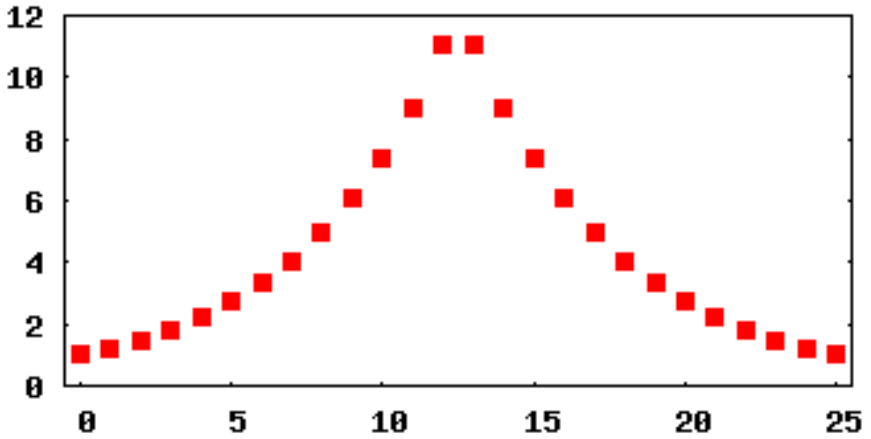
where $d(p,p')$ is the Euclidean distance between the spots of p and p'.

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

Outline

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

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)

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

Previous Work: Partitioning

More recently, a partitioning algorithm was proposed:
Centroid-based Quadrisection (Kahng *et al.*, 2003)

- Recursive partitioning; the chip is divided into four quadrants
- Each sub-problem is treated as a separate placement
  - A placement algorithm is needed in the end
- Reduce run-time; may improve placement

Pivot Partitioning: Motivation

E.Coli GeneChip

Number of probes	%	Numer of embeddings
1 765	0.78	1
28 410	12.63	26
52 913	23.52	351
63 588	28.26	3 276
48 257	21.45	23 751
22 628	10.06	142 506
6 372	2.83	736 281
957	0.43	3 365 856
86	0.04	13 884 156
224 976	100.00	

Observation

- Some probes have only a few possible embeddings
- Others may have up to several millions
- Probes with more embeddings are more "flexible"
  - They can adapt better to their neighbors

Pivot Partitioning: Pivots

- We use probes with fewer embeddings (pivots) to:
  - Drive the partitioning of the probe set
  - Re-embed the probes before their placement

Algorithm 1: PivotPartitioning

Input: chip dimensions, set of probes $\mathcal{P} = \{p_1, p_2, ...p_n\}$, and maximum partitioning depth t_{max}

Output: placement of the probes $p \in \mathcal{P}$ on the chip

1. Select probes p with minimum number of embeddings, $E(p)$:
   a) Let $\mathcal{Q} = \{p \in \mathcal{P} | E(p) \text{ is minimal}\}$
   b) $\mathcal{P} \leftarrow \mathcal{P} \setminus Q$
2. Let R be a region consisting of all spots
3. Return RecursivePartitioning (1, t_{max}, R, $\mathcal{Q}$, $\mathcal{P}$)

Pivot Partitioning: Recursive Partitioning

Algorithm 2: RecursivePartitioning

Input: current depth t, maximum depth t_{max}, region R, pivot candidates $\mathcal{Q}$, non-pivot probes $\mathcal{P}$,
Output: placement of probes $p \in \mathcal{P} \cup \mathcal{Q}$ on R

1. If $t = t_{max}$ then
   a) Re-embed $p \in \mathcal{P}$ optimally with respect to all $q \in \mathcal{Q}$
   b) Return RowEpitaxial (R, $\mathcal{P} \cup \mathcal{Q}$)
2. Select q' and $q'' \in \mathcal{Q}$ with maximum conflicts $c(q', q'')$
3. Partition $\mathcal{P}$ and $\mathcal{Q}$:
   a) Let $\mathcal{Q}' = \{q \in \mathcal{Q} \mid c(q, q') < c(q, q'')\}$; $\mathcal{Q}'' \leftarrow \mathcal{Q} \setminus \mathcal{Q}'$
   b) Let $\mathcal{P}' = \{p \in \mathcal{P} \mid c(p, q') < c(p, q'')\}$; $\mathcal{P}'' \leftarrow \mathcal{P} \setminus \mathcal{P}'$
4. Partition R into R' and R'' proportionally to $\mathcal{P}' \cup \mathcal{Q}' / \mathcal{P}'' \cup \mathcal{Q}''$
5. Return RecursivePartitioning ($t + 1$, t_{max}, R', $\mathcal{Q}'$, $\mathcal{P}$')
   $\cup$ RecursivePartitioning ($t + 1$, t_{max}, R'', $\mathcal{Q}''$, $\mathcal{P}$")

Pivot Partitioning

Similar to the Centroid-based Quadrisection but...

- Alternate horizontal and vertical partitions
- Use probes with fewer embeddings as pivots ("centroids")
  - Faster selection, better representatives
- First algorithm to combine placement and embedding
  - Consider all embeddings when assigning probes to regions
  - Re-embed probes optimally in regards to pivots (with OSPE) before the placement

Pivot Partitioning: Results on Random Chips

Border Length Minimization (cost: normalized border length)

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	42.77	34	39.19	13	40.72	10	42.11	11
200	41.63	429	37.30	155	38.53	62	40.00	85
300	41.38	1 174	36.12	766	37.22	264	38.53	139
500	41.27	3 524	34.69	3 472	35.50	1 996	36.58	713

Conflict Index minimization (cost: average conflict index)

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	514.49	45	453.67	37	467.78	19	475.44	15
200	517.07	192	466.22	215	452.41	166	462.55	99
300	518.51	438	475.84	524	452.00	466	448.17	336
500	517.50	1 471	481.36	1 530	462.33	1 472	445.43	1 295

Pivot Partitioning (PP) × Centroid-based Quadrisection (CQ)

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	393 218	0.18%	399 312	-1.89%	410 608	-2.48%
200	1 524 803	2.27%	1 545 825	0.48%	1 573 096	-1.34%
300	3 493 552	7.12%	3 413 316	2.05%	3 434 964	-0.61%
500	9 546 351	8.95%	9 355 231	4.67%	9 307 510	1.03%

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!

- 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