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

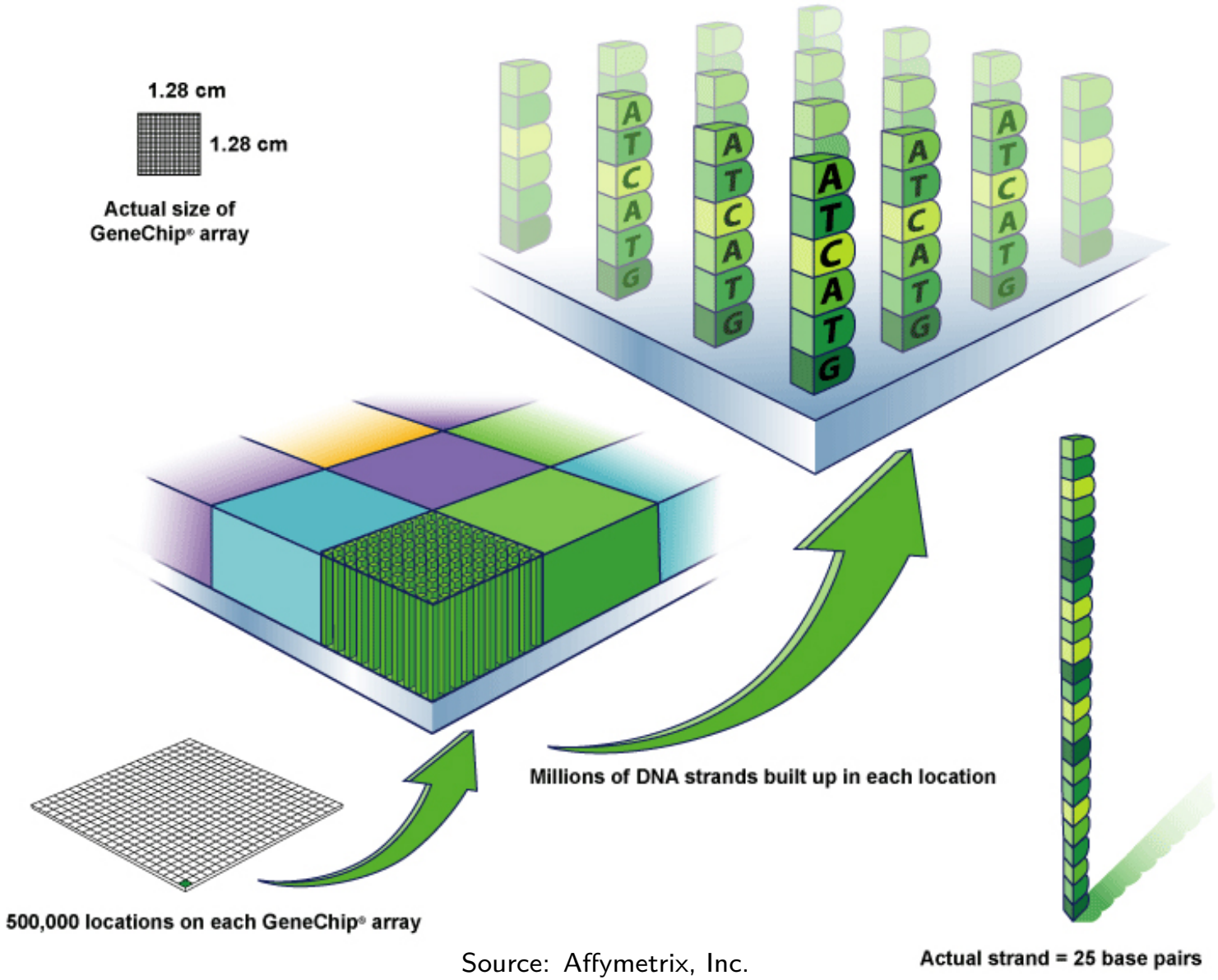
[3]Graduiertenkolleg Bioinformatik

6th Workshop on Algorithms in Bioinformatics
Zürich, Sep 2006

Outline

Outline

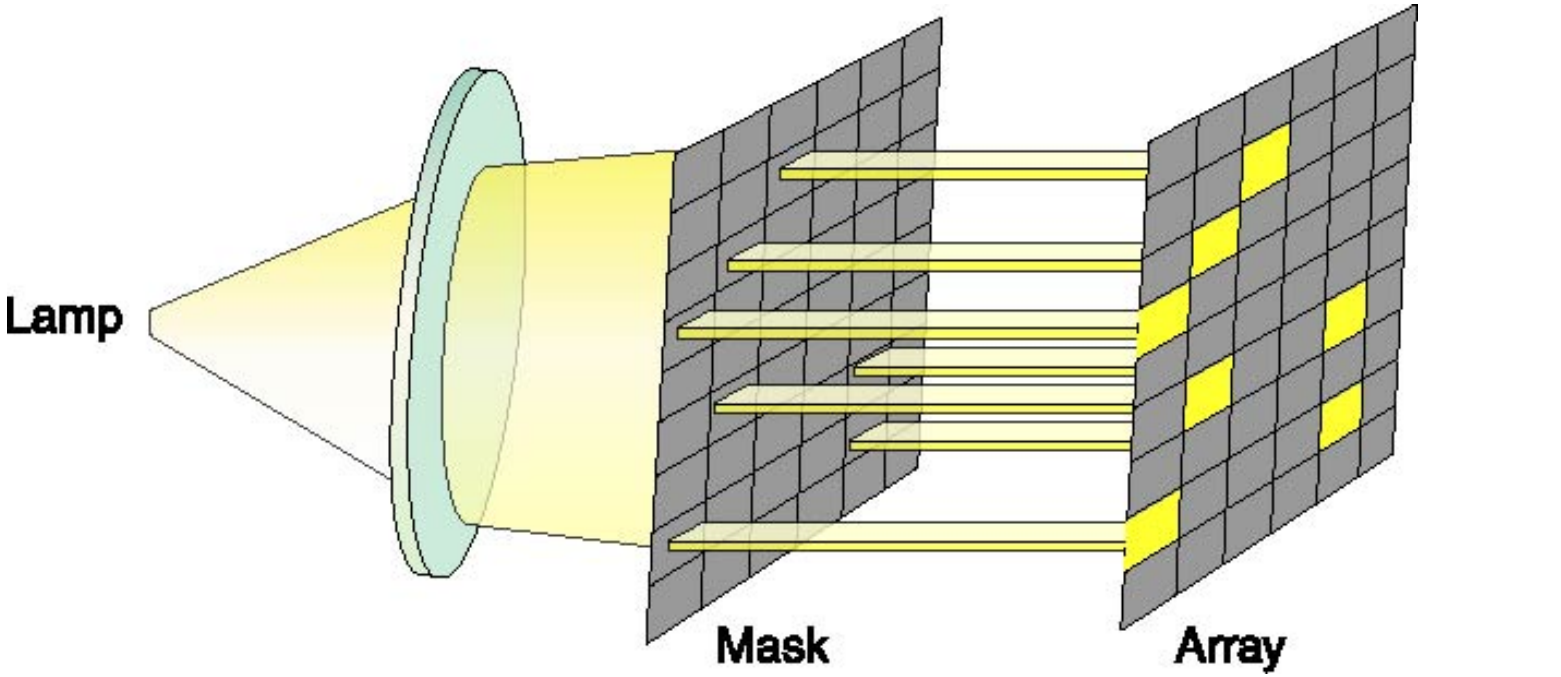
High-Density Oligonucleotide Microarrays



Source: Affymetrix, Inc.

Probe Synthesis: Photolitographic Masks

Source: Affymetrix, Inc.



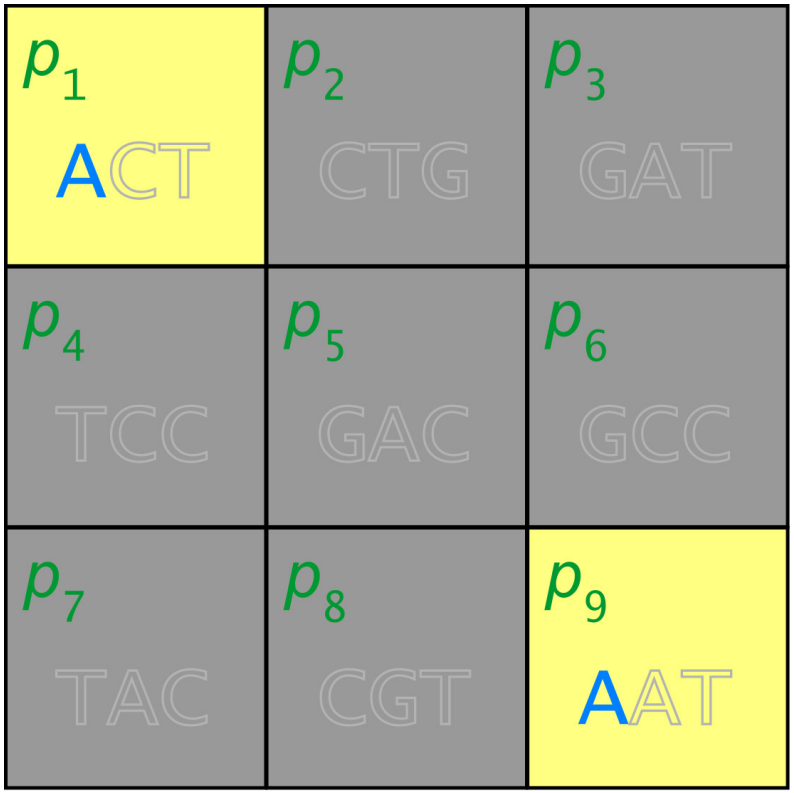
- Probes are synthesized on the chip in a series of steps
- Each step appends a particular nucleotide to selected regions
- Selection occurs by exposure to light

Deposition Sequence and Probe Embeddings

p_1 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

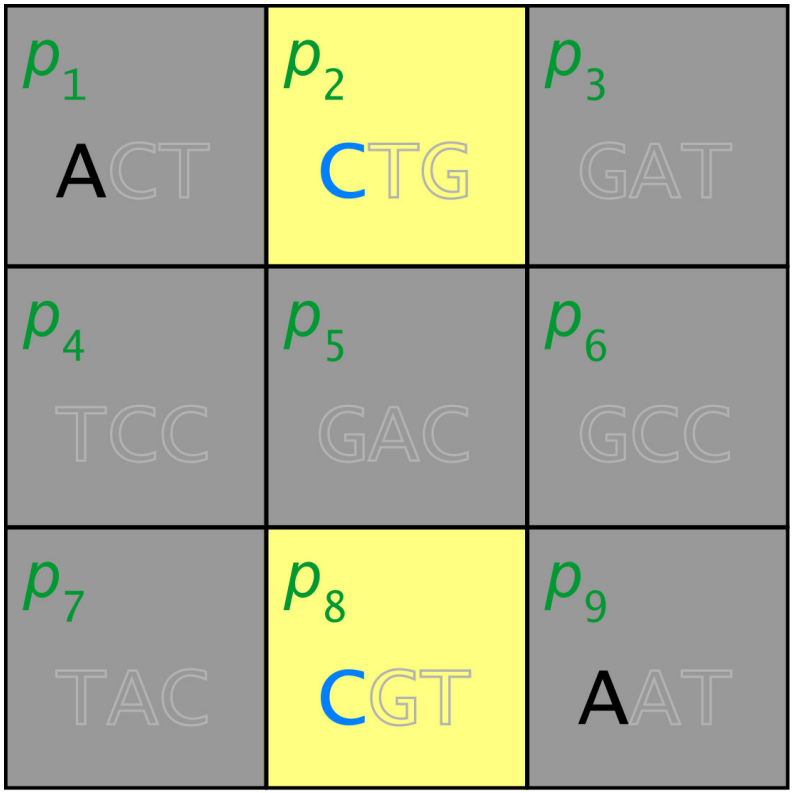
$$
\begin{aligned}
S &= \text{ACGTACGTACGT} \\
\varepsilon_1 &= ------------ \\
\varepsilon_2 &= ------------ \\
\varepsilon_3 &= ------------ \\
\varepsilon_4 &= ------------ \\
\varepsilon_5 &= ------------ \\
\varepsilon_6 &= ------------ \\
\varepsilon_7 &= ------------ \\
\varepsilon_8 &= ------------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

Deposition Sequence and Probe Embeddings



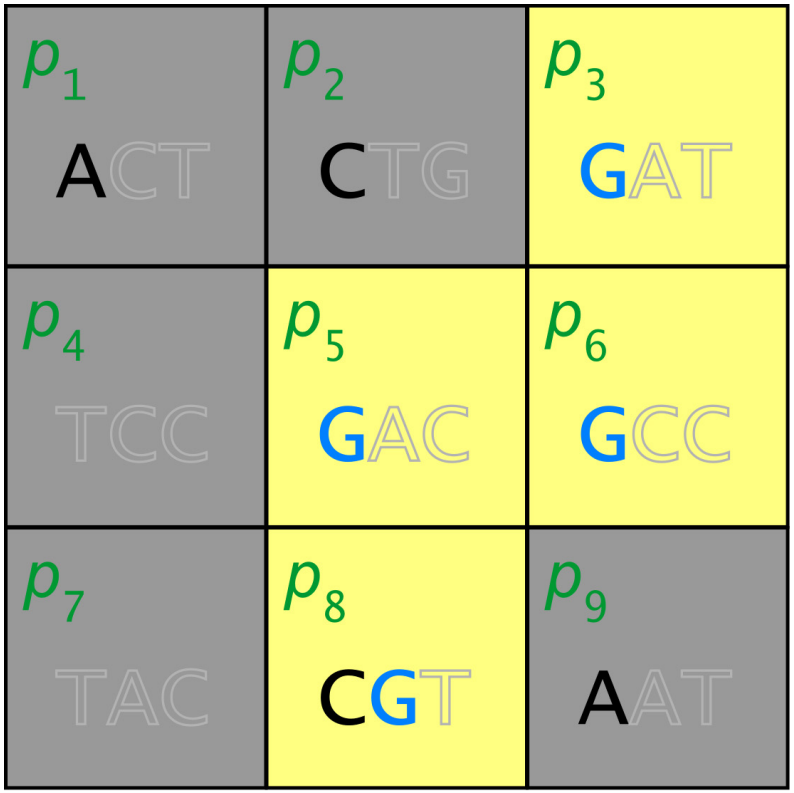
$$
\begin{aligned}
S &= \text{ACGTACGTACGT} \\
\varepsilon_1 &= \text{A}\text{-----------} \\
\varepsilon_2 &= \text{-----------} \\
\varepsilon_3 &= \text{-----------} \\
\varepsilon_4 &= \text{-----------} \\
\varepsilon_5 &= \text{-----------} \\
\varepsilon_6 &= \text{-----------} \\
\varepsilon_7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

Deposition Sequence and Probe Embeddings



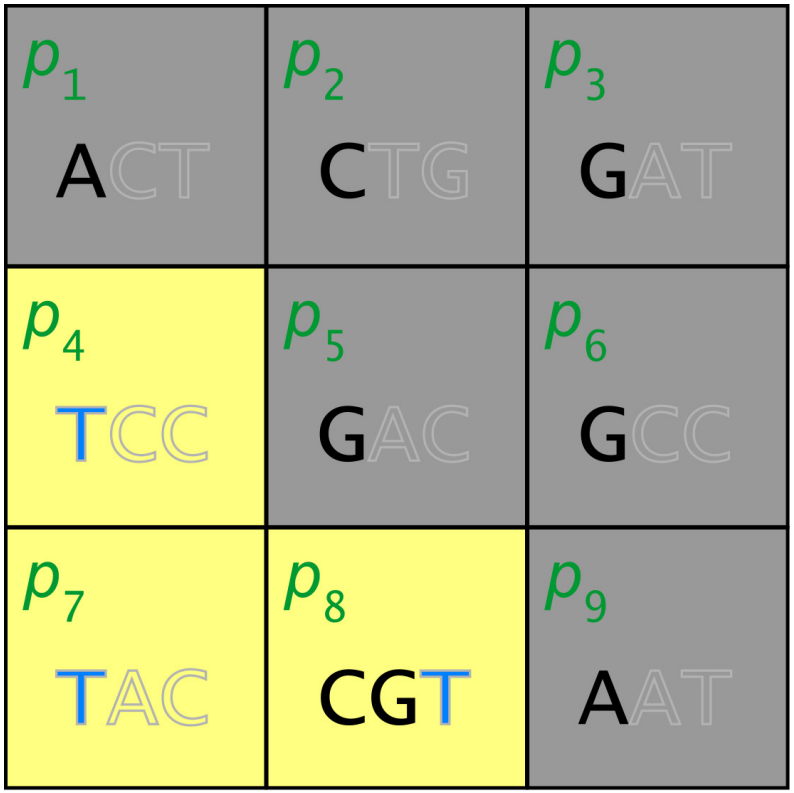
$$
\begin{aligned}
S &= \text{ACGTACGTACGT} \\
\varepsilon_1 &= \text{A----------} \\
\varepsilon_2 &= \text{-C---------} \\
\varepsilon_3 &= \text{----------} \\
\varepsilon_4 &= \text{----------} \\
\varepsilon_5 &= \text{----------} \\
\varepsilon_6 &= \text{----------} \\
\varepsilon_7 &= \text{----------}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

Deposition Sequence and Probe Embeddings



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

Deposition Sequence and Probe Embeddings



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

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

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

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

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

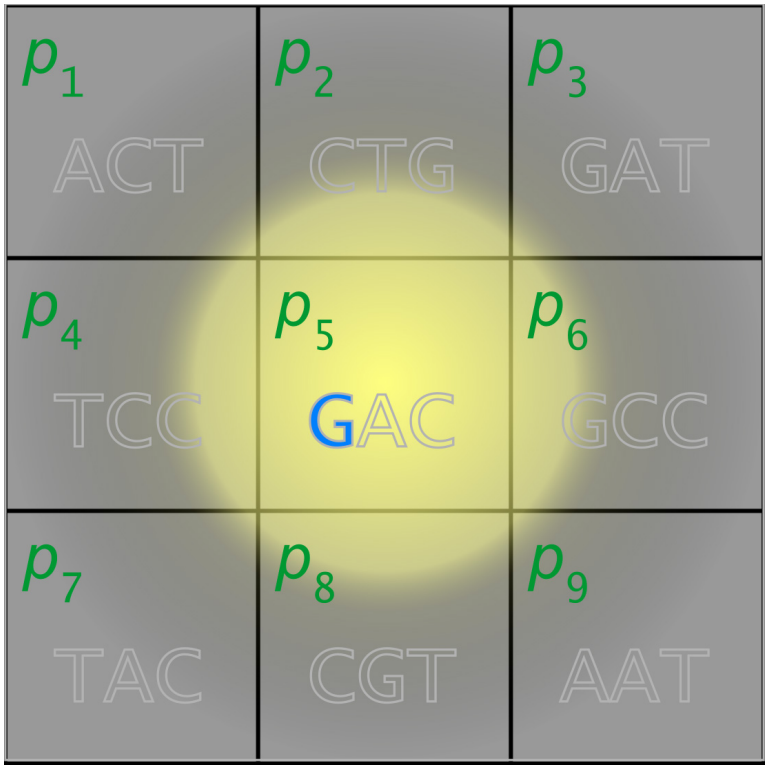
Left-most embedding!

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

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

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

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

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

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 $\delta(p,p',t)$

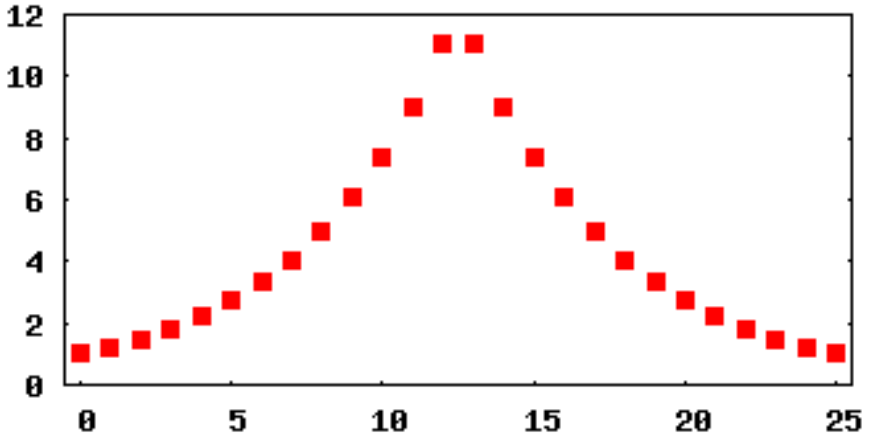
$$\delta(p,p',t) := \begin{cases} (d(p,p'))^{-2} & \text{if } p' \text{ is unmasked at step } t, \\ 0 & \text{otherwise,} \end{cases}$$

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 ℓ_p is the length of probe p.

Outline

Previous work

- The placement problem has been traditionally approached in two phases:
  1) Placement of probes given a fixed embedding
  2) Re-embedding probes given a fixed placement
- Additionally, a partitioning of the chip may be done to divide the problem into smaller sub-problems

Summary

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!