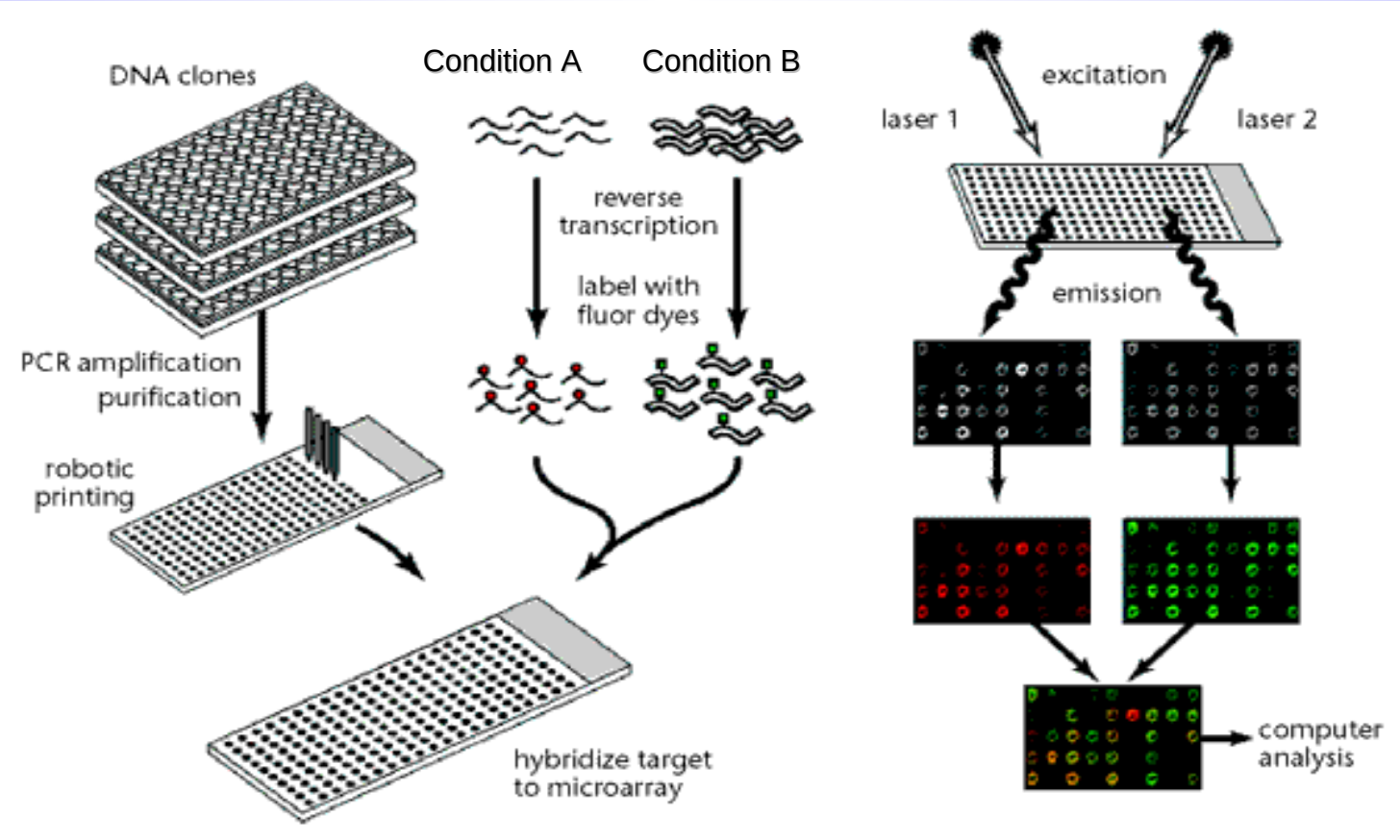
Optimale Versuchsplanung dreifaktorieller cDNA-Microarray-Experimente

1. Einleitung, Modell, Beispiele
2. Optimalität diskreter Designs
3. Ergebnisse
4. Diskussion

Dipl.-Stat. Sven Stanzel
Institut für Medizinische Statistik
RWTH Aachen

GMDS 2006, Leipzig, 11. September 2006

cDNA-Experiment



aus: Duggan et al., Nature Genetics (1999)

Statistisches Modell (1)

2×I×k Design (**2** Farben, **I** Treatments, **k** Zelllinien)

- **Fixed** effects **gene-specific** linear model for **log-ratios** (Landgrebe et al., 2006):

$$Z_t = \delta_g - \delta_r + \tau_{ij} - \tau_{i'j'} + \varepsilon_t;\ \varepsilon_t \sim \left(0; \sigma^2\right) \qquad \textbf{(1)}$$

t = t(g, r, i, i′, j, j′); t = 1,...,N

Z_t: Log Ratio der Fluoreszenzintensitäten auf Array t (für ein spezielles Gen)

δ_g: Farbeffekt des grünen Farbstoffs

δ_r: Farbeffekt des roten Farbstoffs

τ_{ij}: Kombinationseffekt von Treatment i (i = 1,...,I) und Zelllinie j (j = 1,...,k)

$\tau_{i'j'}$: Kombinationseffekt von Treatment i′ (i′ = 1,...,I) und Zelllinie j′ (j′ = 1,...,k)

Statistisches Modell (2)

- **Matrixnotation:** $\mathbf{Z} = \mathbf{X}\theta + \varepsilon$ **(2)**

$\mathbf{Z} = (Z_1,\ldots,Z_N)^T \sim N\times 1$ Vektor der beobachteten Log Ratios

$\theta = (\delta_g,\delta_r,\tau_{11},\ldots,\tau_{lk})^T \sim s\times 1$ Parametervektor (s:=lk+2)

$\mathbf{X} \sim N\times s$ Designmatrix

$\varepsilon = (\varepsilon_1,\ldots,\varepsilon_N)^T \sim (0_N , \sigma^2 I_N)$ Residualvektor

- **Kontraste:**
  - Kontrastmatrix: $\mathbf{K} \sim s\times q$; $Rg(\mathbf{K}) := r$, $r<s$, $r\leq q$ (abh. von K !!)
  - lineare Kontraste: $\mathbf{K}^T\theta$ mit $\mathbf{K}^T\mathbf{1}_s = \mathbf{0}_q$

Beispiel (**l=2** Treatments: 1, 2 ; **k=3** Zelllinien: a, b, c):

K^T:

g	r	a1	a2	b1	b2	c1	c2
0	0	1	-1	1	-1	1	-1

(Treatmenteffekt)

2×2×3 Designs - Beispiele

(Landgrebe et al., 2006)

n = 6 Arrays **; l = 2** Treatments (1, 2); **k = 3** Zelllinien (a, b, c)

Within Treatment Swap (TS)

Designmatrix (konkretes Design):

Array	g	r	a1	a2	b1	b2	c1	c2
1	1	-1	1	0	-1	0	0	0
2	1	-1	0	0	1	0	-1	0
3	1	-1	-1	0	0	0	1	0
4	1	-1	0	1	0	-1	0	0
5	1	-1	0	0	0	1	0	-1
6	1	-1	0	-1	0	0	0	1

Within Cell Line Swap (CLS)

Designmatrix (konkretes Design):

Array	g	r	a1	a2	b1	b2	c1	c2
1	1	-1	1	-1	0	0	0	0
2	1	-1	-1	1	0	0	0	0
3	1	-1	0	0	1	-1	0	0
4	1	-1	0	0	-1	1	0	0
5	1	-1	0	0	0	0	1	-1
6	1	-1	0	0	0	0	-1	1

„E-Effizienz“

(Landgrebe et al., 2006)

- **CLS (2×)** hat **maximale „E-Effizienz“** (unter 9 betrachteten 12-Array-Designs) für das Schätzen von **Treatmenteffekt** sowie **Interaktionseffekt** von Treatment und Zelllinie.
- **TS (2×)** hat **maximale „E-Effizienz“** (unter 9 betrachteten 12-Array-Designs) für das Schätzen des **Zelllinieneffektes.**
- **Schwachstellen:**
  - (globale) Optimalität nicht gesichert, da keine wirkliche E-Effizienz
  - stattdessen: Vergleich ausgewählter Designs anhand des E-Kriteriums
  - Effizienzberechnungen abhängig von Anzahl Arrays (→ wenig flexibel !)
  - wenige, ausgewählte Designs in Effizienzvergleich einbezogen (**lokale** Optima)
  - Abhängigkeit vom Optimalitätskriterium (z.B. D-, A-Kriterium) nicht untersucht
- **Ziele:** 1) Herleitung **global** optimaler Versuchspläne (+ **analytischer** Beweis!)
  2) **flexible** Anzahl von N, I, k
  3) Untersuchung der **Robustheit** gegenüber **Kriterium**

Äquivalenztheorem (Φ_p-Optimalität)

- $m = 2 \cdot \binom{lk}{2}$: Anzahl theoretisch möglicher verschiedener Designpunkte $x_1,\ldots,x_m$
- $\chi = \{\mathbf{x}_1,\ldots,\mathbf{x}_m\}$: Designraum (diskrete Menge von Designpunkten)
- Ω: Menge aller möglichen **diskreten** Designs $\xi = \begin{Bmatrix} x_1,\ldots,x_m \\ p_1,\ldots,p_m \end{Bmatrix}; 0 \le p_i \le 1 \; \forall i = 1,\ldots,m; \sum_{i=1}^{m} p_i = 1$
- $\mathbf{M} := \sum_{i=1}^{m} p_i \mathbf{x}_i \mathbf{x}_i^T$ (~ s×s) **Momentenmatrix**
- ξ^***:** Kandidatendesign; *schätzbar* bzgl. $\mathbf{K}^T\theta$

Theorem 1 (Äquivalenztheorem für Matrizenmittelwerte, *generalisiert*):

(a) ξ^* ist ϕ_p**-optimal** für $\mathbf{K}^T\theta$ dann und nur dann, wenn eine *g-Inverse* **G** von **M** existiert, die die **Normalitätsungleichung**

$$\mathbf{x}^T\mathbf{GK}(\mathbf{K}^T\mathbf{GK})^+(\mathbf{K}^T\mathbf{GK})^{1-p}(\mathbf{K}^T\mathbf{GK})^+\mathbf{K}^T\mathbf{G}^T\mathbf{x} \le Tr\left[(\mathbf{K}^T\mathbf{GK})^+(\mathbf{K}^T\mathbf{GK})^{1-p}\right] \qquad \textbf{(N)}$$

für **alle** Designpunkte $x \in \chi$ erfüllt (für alle $p \in (-\infty;1]$).

(b) Für die **Trägerpunkte** von ξ^* gilt **Gleichheit** in **(N).**

(vgl. Pukelsheim, 1993, p. 180/205)

Within Cell Line Swap Design **(CLS)**

Trägerpunkte vergleichen jeweils **zwei Treatments** bei Verwendung **derselben Zelllinie.**

Theorem 2 (Optimalität des Within Cell Line Swap Designs):

Das Desígn **CLS** ist

(a) D-optimal (p=0) (b) A-optimal (p=-1)

(c) E-optimal (p=-∞)

für das Schätzen der folgenden (linearen) Kontraste:

(1) **Treatmenteffekt** (für alle k, l ∈ N)

(2) **Interaktionseffekt** von Treatment und Zelllinie **(für k ≥ l ;** k, l ∈ N)

(3) **Kombination** von **Treatmenteffekt** und **Interaktionseffekt** (für alle k, l ∈ N)

Beweisskizze (Treatmenteffekt; ϕ_p)

Design CLS: $$\mathbf{M} = c_1 \cdot \begin{bmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{I_k} \otimes \left(\mathbf{A_l}^T \mathbf{A_l}\right) \end{bmatrix} \quad , \mathbf{G} = \frac{1}{c_1} \cdot \begin{bmatrix} \mathbf{A}^- & \mathbf{0} \\ \mathbf{0} & \mathbf{I_k} \otimes \left(\frac{1}{4l^2} \mathbf{A_l}^T \mathbf{A_l}\right) \end{bmatrix} \quad , \mathbf{A_l} = \begin{bmatrix} \tilde{\mathbf{A}}_\mathbf{l} \\ -\tilde{\mathbf{A}}_\mathbf{l} \end{bmatrix} \quad .$$

Kontrastmatrix: $$\mathbf{K} = \left[\mathbf{0} \quad \mathbf{0} \quad \mathbf{1_k}^T \otimes \tilde{\mathbf{A}}_\mathbf{l}\right], \; \tilde{\mathbf{A}}_\mathbf{l} \sim \binom{l}{2} \times l \quad .$$

$$\mathbf{K}^T\mathbf{G}^T = \frac{1}{c_1} \cdot \left[\mathbf{0} \quad \mathbf{1_k}^T \otimes \frac{1}{2l} \tilde{\mathbf{A}}_\mathbf{l} \right] \quad , \; \mathbf{K}^T\mathbf{G}\mathbf{K} = c_2 \cdot \tilde{\mathbf{A}}_\mathbf{l} \tilde{\mathbf{A}}_\mathbf{l}^T \quad , \; \left(\mathbf{K}^T\mathbf{G}\mathbf{K}\right)^+ = c_3 \cdot \tilde{\mathbf{A}}_\mathbf{l} \tilde{\mathbf{A}}_\mathbf{l}^T$$

Designpunkte von CLS: $\mathbf{x} = [1, -1, x_{11}, \ldots, x_{lk}]^T$ mit

$$x_{uv} = \left\{ \begin{array}{l} +1, \text{ Treatment u} \leftrightarrow \text{Zelllinie v (g)} \\ -1, \text{ Treatment u} \leftrightarrow \text{Zelllinie v (r)} \\ 0, \text{ sonst} \end{array} \right\}; \; u = 1, \ldots, l; \; v = 1, \ldots, k \quad \text{und} \quad \bar{x}_{\cdot w} = \frac{1}{l} \sum_{u=1}^{l} x_{uw}.$$

Dann gilt: $$\mathbf{x}^T \mathbf{G}\mathbf{K} \left(\mathbf{K}^T\mathbf{G}\mathbf{K}\right)^+ \left(\mathbf{K}^T\mathbf{G}\mathbf{K}\right)^{1-p} \left(\mathbf{K}^T\mathbf{G}\mathbf{K}\right)^+ \mathbf{K}^T\mathbf{G}^T\mathbf{x} \le Tr\left[\left(\mathbf{K}^T\mathbf{G}\mathbf{K}\right)^+ \left(\mathbf{K}^T\mathbf{G}\mathbf{K}\right)^{1-p}\right]$$

$$\Leftrightarrow \sum_{u=1}^{l} \sum_{v=1}^{k} \left[x_{uv} \cdot \sum_{w=1}^{k} \left(x_{uw} - \bar{x}_{\cdot w}\right) \right] \le 2. \quad (\Delta)$$

Gültigkeit von (Δ) für alle x (Gleichheit für Trägerpunkte von CLS) kann über *Fallunterscheidung* gezeigt werden.

■

Within Treatment Swap Design **(TS)**

Trägerpunkte vergleichen jeweils **zwei Zelllinien** bei Verwendung **desselben Treatment.**

Theorem 3 (Optimalität des Within Treatment Swap Designs)

Das Desígn **TS** ist für **l ≥ k**

(a) D-optimal (p=0)

(b) A-optimal (p=-1)

(c) E-optimal (p=-∞)

für das Schätzen des linearen Kontrastes für den **Interaktionseffekt** von Treatment und Zelllinie.

Beweis: Analog zum Beweis von Theorem 2. ■

Diskussion

- Für zahlreiche *dreifaktorielle* Designsituationen konnten **global optimale** cDNA-Microarray-Designs gefunden und deren Optimalität **analytisch** bewiesen werden.
- **Robustheit:** Wahl des optimalen Designs **unabhängig** vom **Designkriterium** !!
- N, I, k nicht auf bestimmte Werte beschränkt („hohe Flexibilität“).
- In manchen Situationen ist die Lösung abhängig von der *Zahl der Zelllinien* bzw. der *Zahl der Treatments*.
- **Praxis:** - Empfehlungen für Wahl des effizientesten cDNA-Microarray-Designs
  - Möglichkeit der Einsparung finanzieller Ressourcen

Ausblick:

- Beweis der Optimalität von **TS** bzgl. Zelllinieneffekt
- Betrachtung **anderer Kontraste** (z.B. Helmert-Kontraste)
- Betrachtung **vier- und mehrfaktorieller** Designsituationen
- Optimales Design **modellabhängig** ??
- Berücksichtigung von **Korrelationsstrukturen** zwischen Genen

Literatur

Duggan, D., Bittner, M., Chen, Y., Meltzer, P., Trent, J. (1999): *Expression profiling using cdna Microarrays.* Nature Genetics 21 (1 Suppl): 10-14.

Landgrebe, J., Bretz, F., Brunner, E. (2006): *Efficient design and analysis of two colour factorial microarray experiments.* Computational Statistics and Data Analysis 2006; 50: 499-517.

Pukelsheim, F. (1993): *Optimal Design of Experiments*. Wiley, New York.

Searle, S. R. (1971): *Linear Models*. Wiley, New York.

Searle, S. R. (1982): *Matrix Algebra useful for Statistics.* Wiley, New York.