Microarray hybridization: 
Learning about RNA-degradation, G-stacks and binding isotherms

© Hans Binder & Mario Fasold
University Leipzig
Interdisciplinary Centre for Bioinformatics

Plön May/2011
Microarrays are tools for studying biology (expression, copy numbers, genotypes...)

Calibration: extracting proper expression values

Quality control
Microarrays are tools for studying

1. biology
   (expression, copy numbers, genotypes...)

   **Calibration:**
   extracting proper expression values

2. surface hybridization
   (isotherms, molecular interactions)

   **Surface hybridization:**
   extracting rules and parameters

   \[ I = f(c, \text{conditions}) \]

   \[ I = f(\text{sequence, cond}) \]
Microarrays are tools for studying biology.

**Calibration:**
extracting proper expression values

**Surface hybridization:**
extracting rules and parameters

- $I = f(c, \text{conditions})$
- $I = f(\text{sequence, cond})$
Bad and good things about microarrays

a) bad things about microarrays – effects which distort intensity
- non-specific background
- sequence-specific affinity
- saturation of probe spots
- RNA degradation, 3'/5'-bias
- specific sequence motifs: e.g. GGG

Good things:
- To solve: Suited rules (isotherms, interactions) and sufficient set-parameters that can be estimated for each array
- Physico-chemical stuff: e.g. degradation and GGG-effects
The inverse (Langmuir-) isotherm problem

\[ I \sim \frac{M \times X}{1 + X} \]

\[ X = \frac{[S]}{c_{50\%}} + N \]

Irvin Langmuir (1881-1957)
Nobel prize 1932
The inverse (Langmuir-) isotherm problem

Options to estimate the parameters:

1. Special calibration data (e.g. spiked-in) for each array

2. Use 'ordinary' probe data as intrinsic reference ('hook' method)
   → Expression values (inverse Langmuir solution)
   → Quality control

N-background: N

target concentration, [S]
Situation 1: PM/MM-pairings

targets

identical targets

strong affinity: $c_{50\%}^{\text{probe}}$

weak-affinity: $c_{50\%}^{\text{ref}}$

probes

internal reference

microarray

PM

MM

allele A

cross allele B

\text{'old affy' expression}

SNP-arrays
Situation 1: PM/MM-pairings

targets

strong affinity: \( c_{50\%}^{\text{probe}} \)
weak affinity: \( c_{50\%}^{\text{ref}} \)

probes
internal reference
microarray

PM
allele A
Matched

MM
allele B
mismatched duplexes

'sold affy' expression

SNP-arrays
Situation 2: PM-only

- Targets interrogating the same transcripts

- Strong affinity: $c_{50}^{\text{probe}}$
- Weak affinity: $c_{50}^{\text{ref}}$

- probes
- internal reference
- microarray

- PM(strong)
- PM(weak)
- PMonly expression

- Strongly-bound
- weakly-bound
- matched duplexes
Situation 2: PM/PM-pairings

**targets**

- **strong affinity**: $c_{50%}^{probe}$
- **weak affinity**: $c_{50%}^{ref}$

**probes**

- PM(strong)
- PM(weak)

**internal reference**

- PMonly expression

**microarray**

- Strongly-bound
- weakly-bound
- matched duplexes
Situation 3: high/low-conc-pairings

Targets interrogating the same transcripts

High conc.: \([S]^{\text{high}}\)

Low conc.: \([S]^{\text{low}}\)

probes

internal reference

microarray

\begin{align*}
PM(3') & \quad PM(5') & \quad PM\text{only expression} \\
PM(\text{before wash}) & \quad PM(\text{after wash})
\end{align*}

Strongly-bound  \quad weakly-bound  \quad matched duplexes
Situation 3: high-conc/low-conc-pairings

Targets interrogating the same transcripts

High conc.: \([S]^{\text{high}}\)

Low conc.: \([S]^{\text{low}}\)

probes

internal reference

microarray

\begin{align*}
\text{PM}(3') & \quad \text{PM}(5') \\
\text{PM(before wash)} & \quad \text{PM(after wash)}
\end{align*}

Highly-weakly populated duplexes
General: high/low-binding strength

Targets interrogating the same transcripts

High: $X$

Low: $X^{\text{ref}}$

$X = [S]/c_{50\%}$

Highly- weakly- populated duplexes

probes internal reference microarray
Example: PM/MM-pairings

Let's use the MM as internal reference, by calculating

- the difference

$$\Delta \equiv \log PM - \log MM$$

- the mean

$$\Sigma \equiv \frac{1}{2}(\log PM + \log MM)$$
The $\Delta$-vs-$\Sigma$ plot

... and by plotting $\Delta$-vs-$\Sigma$ into a sort of M-A-plot
The hybridization regimes

... can be clearly identified: N... non-specific binding
The hybridization regimes

... can be clearly identified:

mix... N- and S-hybridization
The hybridization regimes

... can be clearly identified:

S... specific hybridization
The hybridization regimes

... can be clearly identified:

sat... saturation
The hybridization regimes can be clearly identified: as asymptotic
Position and dimensions of the hook curve

regime: \[\text{N | mix | S | sat | as}\]

\[\Delta \quad \Delta_{\text{start}} \quad \Sigma_{\text{start}} \quad \Sigma\]

\[N\text{-intensity} \quad M\text{-}(\text{sat})\text{-intensity}\]

\[\beta \sim -\log X^N\]

\[\alpha \sim \log (X^s / X^s,\text{ref})\]
Mean $\Delta$-vs-$\Sigma$ plot of the chip

We consider all probe-pairs of a chip
- we calculate the average $\Delta$-vs-$\Sigma$ trajectory by applying "brutal" smoothing
- provides the mean isotherm of a particular hybridization

...all probe intensities are corrected for the optical background using the Affymetrix zone-algorithm
Hook analysis: overview

raw intensities → Δ-vs-Σ plot → raw hook

chip characteristics, expression measures → corrected hook → sensitivity profiles

positional-dependent NN model
the (raw) hook – curve reflects the...

Hybridization regimes:

N... non-specific hybridization dominates

mix... N+S hybridization

S... specific binding dominates

sat... progressive saturation

(as... asymptotic range, the probes are saturated)

...which we know from the single probes
We suggest:
the initial break can be used
to identify „absent“ probes,
i.e. which predominantly
hybridize non-specifically
because their specific transscritps are virtually absent
in the hybridization solution

slope(N) ~ correlations PM/MM
slope(mix) ~ S-dependence of the isotherm

break
Fit of the hook-equation provides…

- **height**, $\alpha$ (PM/MM-affinity gain)
- **Width**, $\beta$ (N-strength)
- **Maximum intensity**, $M$

![Graph showing fit parameters](image)
Chip generations (human genome)

- **fit**
- **corrected**
- **raw**

Spot size: 20 μM

# probe sets: 16,000

Improvements:
- probe selection: genomic, thermodynamic criteria
- hybridization, labeling kits, scanner etc.
Hook-pattern I: shift

changes of the sensitivity-settings (scanner, amount of labelling) shift the whole curve in horizontal direction
Hook pattern II: widening

dilution (less total RNA) increases the width of the hook
Hook pattern III: vertical stretch

SNP-array

Probes: P-A, P-B

 targets: only A (homozygote)

PM/MM-gain modifies the height of the hook
Hook pattern IV: „blow up“

washing increases the width AND the heigth of the hook
Simple rules reveal underlying effects…

- Scanner settings
- Dilution
- MM design
- Washing
Clinical trial: Burkitts Lymphoma: 221 patients

Scanner sensitivity

\[ \alpha, \beta, \log M, \log N, \text{abs}, 3'/5'-\text{bias} \]

probe # ~ time (3 years), laboratories, sample:

Scanner settings

IZBI University of Liège
Dilution (less RNA (30%))
degradation hook

Depletion of targets with increasing distance from 3'-end defines the height of the degradation hook.
degradation hook: example

PM/MM-hook

degradation-hook

\[
\Delta, \Delta \Sigma_{3'/5'} \quad \Delta \gamma_{3'/5'} = 0.57 \\
\beta = 2.3 \\
\Sigma(0) = 2.1 \\
M = 4.4
\]

bad RNA

\[
\Delta, \Delta \Sigma_{3'/5'} \quad \Delta \gamma_{3'/5'} = 0.29 \\
\beta = 2.1 \\
\Sigma(0) = 2.2 \\
M = 4.3
\]

good RNA
Degradation threshold

Threshold hook

Human tissue data

Positive Predictive Value = \frac{TP}{TP + FP} = 0.5

Specificity = \frac{TN}{TN + FP} = 0.8
Thank you!

Papers see

http://www.izbi.uni-leipzig.de/izbi/mitarbeiter/Binder/binderpubl.php