

# **Analyse von Microarray Genexpressionsdaten**

Yusuf Tanrikulu      Ewgenij Proschak

15. September 2004

# Inhaltsverzeichnis

<b>1</b>	<b>Einleitung</b>	<b>3</b>
<b>2</b>	<b>Microarrays</b>	<b>4</b>
<b>3</b>	<b>Batcheffekte und Visualisierungstechniken</b>	<b>5</b>
3.1	Batcheffekte . . . . .	5
3.2	Visualisierungstechniken . . . . .	6
<b>4</b>	<b>Kalibrierung</b>	<b>9</b>
4.1	Modell nach Chen . . . . .	9
4.2	Modell nach Huber . . . . .	10
<b>5</b>	<b>Parameterabschätzung</b>	<b>13</b>
5.1	Maximum-Likelihood Abschätzer . . . . .	13
5.2	LTS Regression . . . . .	14
<b>6</b>	<b>Mustererkennung</b>	<b>16</b>
6.1	Projektionsmethoden . . . . .	16
6.2	Clusteralgorithmen . . . . .	16
<b>7</b>	<b>Ergebnisse</b>	<b>18</b>
<b>8</b>	<b>Literaturverzeichnis</b>	<b>19</b>

# 1 Einleitung

Um globale Zusammenhänge innerhalb des Genoms nachvollziehen zu können, werden Microarray-Experimente durchgeführt. Microarrays basieren auf der Messung der Genexpressionsraten einer großen Anzahl von Genen eines Individuums oder eines Gewebetyps. Eine mögliche Fragestellung dieser Experimente ist die Analyse von Ähnlichkeiten und/oder Differenzen der Genexpressionsraten zwischen verschiedenen Gewebetypen oder unter unterschiedlichen biologischen Bedingungen. Ein breites Anwendungsfeld der Microarray-Experimente bildet daher die Tumor-Forschung.

Die Datenmengen solcher Experimente sind so groß, dass sich eine manuelle Verarbeitung als sehr schwierig erweist. Aus diesem Grund ist die Zuhilfenahme von rechnergestützten, bioinformatischen Methoden eine Notwendigkeit, die eine effiziente Analyse und Visualisierung der Datenmengen ermöglicht.

Die Forschungsgruppe um Dr. Wolfgang Huber, welcher am Deutschen-Krebs-Forschungsinstitut in Heidelberg tätig ist, verfasste eine Übersicht der aktuellen Konzepte und Methoden [1] auf diesem Gebiet, die im Folgenden vorgestellt werden.

Kapitel 2 beinhaltet die Techniken der Microarray-Experimente, wonach das Auslesen der Daten und Visualisierungsmöglichkeiten in Kapitel 3 beschrieben werden. Danach folgen im Kapitel 4 die eigens von Huber et al. entwickelte Methode zur Kalibrierung der Daten und Parameterabschätzung im Kapitel 5, die für die Transformation der Daten notwendig ist. Die Identifizierung unterschiedlich exprimierter Gene (Kapitel 6) und Mustererkennung (Kapitel 7) bilden den Abschluss der Übersicht.

## 2 Microarrays

Die Microarray-Technologie beruht auf der Hybridisierung komplementärer Nukleinsäurestränge. Dazu werden spezifische einsträngige Nukleotidsequenzen auf einer festen Unterlage in Form eines Arrays immobilisiert. Jede Arrayzelle mit unterschiedlicher Sequenz wird Spot genannt. Als Unterlage dient je nach Anwendung eine Glasplatte, eine Siliziumscheibe oder eine Nylonmembran. Aus einer Zell- /Gewebeprobe werden die Nukleinsäuren extrahiert und mit einem Fluoreszenzfarbstoff oder radioaktivem Marker versehen. Nun wird das Array mit der aufbereiteten Probe behandelt.

Die Menge an Markersubstanz an jedem Spot korreliert mit der Menge des entsprechenden RNA-Transkripts in der Probe. Heutzutage verwendet man üblicherweise zwei Markierungsmethoden, entweder radioaktiv oder durch Fluoreszenz. Die radioaktive Methode wird zusammen mit der Hybridisierung auf Nylonmembranen benutzt. Bei Anwendung der Arrays auf einer Glasscheibe verwendet man fluoreszierende Marker. Ein Spezialfall der letzten Methode ist die sogenannte Stanford-Technologie, bei der dasselbe Array mit zwei Proben mit komplementärer Farbfluoreszenzmarkierung beschickt werden. Es ist nicht möglich, Aussagen über die absolute Menge des RNA-Transkripts zu machen, weil die zahlreichen chemischen Reaktionen, die der Aufbereitung der Probe dienen, zu viele Nebeneffekte auf die Hybridisierung der Nukleotide ausüben. Daher müssen die Ergebnisse innerhalb eines Experiments im Verhältnis zueinander betrachtet werden. Man spricht dabei von Genexpressionsraten. Die Hybridisierung einer Proben-RNA erfolgt meist nur an einem bestimmten Spot, weil die dortige repräsentative DNA-Sequenz entweder ein speziell ausgewähltes cDNA-Fragment ist oder eine Menge von Oligonukleotiden dieses speziellen Gens beinhaltet. Es ist von außerordentlicher Wichtigkeit eine genaue Planung des Microarray-Experiments inklusive des Designs und der Zusammensetzung der Spots und Proben durchzuführen, weil schon kleinste Fehler einen erheblichen Verlust des Genauigkeitsgrades nach sich ziehen und somit die Ergebnisse verfälschen. Besonders interessant ist der Vergleich der Expression eines Gens zu verschiedenen Konditionen. Es kann sich dabei um einen zeitlichen Verlauf oder um unterschiedliche biologische Bedingungen, wie Temperatur, Nahrungsangebot oder Zellzyklusstadium u.v.m., handeln.

# 3 Batcheffekte und Visualisierungstechniken

Das Auslesen der Daten ist der Punkt in dem Experiment, wo die Labortätigkeit aufhört und die statistische Auswertung mit einem Computer beginnt.

Um die Intensität der Genexpressionen auszulesen, wird mit einem Detektor, der eine hohe Bildauflösung besitzt, gescannt. Ziel ist es, jeden Spot des Arrays durch eine möglichst hohe Anzahl von Bildpunkten darzustellen, damit aus diesen Daten ein Messwert für die Fluoreszenzintensität errechnet werden kann. Zusätzlich ist es möglich, bereits in diesem Schritt auch das Hintergrundrauschen der Messung und die Spotqualität zu messen. Je nach Art des Experiments variiert die Anwendbarkeit des einen oder des anderen Auslesealgorithmus. Die Auswahl der geeigneten Methode hängt stark von den verwendeten Techniken ab, wie zum Beispiel:

- die Unterlage: Glas oder Nylon.
- den Pippetiertechnik: Quill-Pen, Pin & Ring oder Ink Jetting.
- die Markierungsmethode: fluoreszent oder radioaktiv.

Im folgenden wird beschrieben welchen Ursprung die Fehler haben können und wie sie frühestmöglich durch geeignete Visualisierungstechniken erkannt werden können.

## 3.1 Batcheffekte

Die Microarray-Experimente sind höchst anfällig für die sogenannten Batcheffekte. Am weitest verbreitet sind folgende Verfälschungen:

**Spotting** Die Menge der Probe in den Nadeln des Roboters, der damit das Array behandelt, kann leicht variieren.

**PCR Amplifikation** Proben, die durch die Polymerase-Kettenreaktion (PCR) erzeugt werden, enthalten oft nicht die gleichen Vielfachen einer Sequenz, da die Amplifikation der unterschiedlichen Nukleotidstränge mit unterschiedlicher Geschwindigkeit verlaufen kann.

**Probenaufbereitung** Bei der Vorbereitung der Proben ist eine Vielzahl komplexer biochemischer Reaktionen, wie zum Beispiel die reverse Transkription, durchzuführen. Diese können von Labor zu Labor und innerhalb eines Experiments Unterschiede aufweisen.

**RNA-Abbau** Unterschiedliche RNA-Stränge haben aufgrund ihrer Sekundärstruktur eine unterschiedliche Halbwertszeit. Um sie zu stabilisieren, werden eine Vielzahl von Gegenmaßnahmen angewendet, die auch Nebeneffekte nach sich ziehen können.

**Array-Beschichtung** Sowohl die Effizienz der Probenfixierung auf dem Array, als auch die Intensität des Hintergrundrauschens hängt stark von der Array-Beschichtung mit der Probe ab.

Diese Probleme sollten beim Design eines Microarray-Experiments beachtet werden. Kann man trotz allem einen Fehler nicht verhindern, so sollten die experimentellen Bedingungen so gewählt werden, dass die biologische Fragestellung nicht beeinflusst wird. Falls zum Beispiel ein Vergleich zwischen zwei Tumorproben durchgeführt werden soll, so ist es ratsam, beide Proben nicht in verschiedenen Labors aufbereiten zu lassen.

## 3.2 Visualisierungstechniken

Die einzelnen Gewebeproben haben üblicherweise die Eigenschaft, dass die meisten Gene im Verlauf des Microarray-Experiments gleich stark exprimiert werden. Dies kann anhand eines Scatterplots für jedes Paar von Proben visualisiert werden, wie es in Abbildung 3.1 gezeigt ist.

Diese Art der Visualisierung ermöglicht es, Fehler, die im Verlauf des Experiments entstanden sind, hervorzuheben. Im Idealfall sollte die Mehrheit der Datenpunkte auf der Winkelhalbierenden des Plots liegen, weil die meisten Gene gleich stark exprimiert werden. Aber in Wirklichkeit ist meistens eine grobe Näherung an den Idealzustand zu beobachten (siehe Abb. 3.1). In dem Beispielplot erkennt man auch, dass die meisten Daten in dem linken unteren Bereich zu finden sind, was eine genauere Analyse dieser Daten nicht ermöglicht. Eine Abhilfe dagegen schafft das Auftragen der Daten auf einer doppel-logarithmischen Skala, welches in Abbildung 3.2 gezeigt ist.

Um eine Drehung um 45 Grad im Uhrzeigersinn zu erreichen, werden die Daten auf die Variablen  $A = (\log R + c) + \log G$  und  $M = (\log R + c) - \log G$  transformiert, wobei R und G die jeweiligen Intensitäten im roten und grünen Kanal repräsentieren. Die Konstante  $c$  wird zu den Intensitäten des roten Kanals addiert, weil dieser generell schwächer ist als die grünen Intensitäten. Abb. 3.2 zeigt außerdem, dass die Varianz in den Genexpressionen nicht konstant ist. Diese Eigenschaft nennt man Heteroskedastizität.

Im nächsten Kapitel wird gezeigt, dass für weitere statistische Analysen die Konstanz der Varianz (Homoskedastizität) in Abhängigkeit vom Mittelwert von äußerster Wichtigkeit ist, weshalb eine Varianzstabilisierende Transformation der Daten vorgenommen wird.

Eine Eigenschaft der Daten ist die typische Verteilung der Intensitäten. Da viele experimentelle Probleme auf dem ganzen Array oder der Samplevorbereitung auftauchen, ist es empfehlenswert das Intensitätenhistogramm jedes Samples anzuschauen. Typischerweise hat ein Array mit zufällig ausgewählten Genen eine unimodale Verteilung (ein Maximum mit einem Abklingen nach rechts, Abbildung 3.3).

In vielen Fällen deuten mehrere Peaks auf ein Experimentierfehler hin, daher ist es ratsam, weitere Visualisierungen der Daten zu betrachten, um Fehler erkennen zu können.

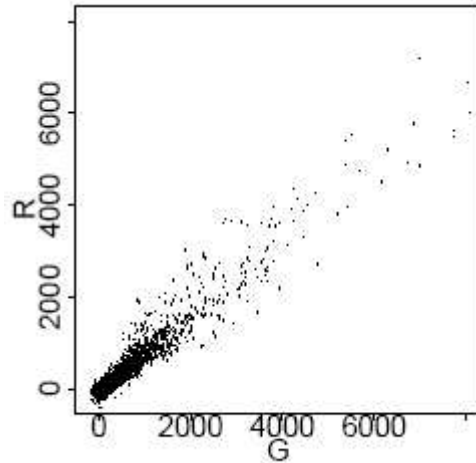


Abbildung 3.1: Ein Scatterplot eines Stanford-MA-Experiments, in dem die Intensitäten des roten und grünen Kanals in willkürlichen Einheiten gegeneinander aufgetragen sind [1].

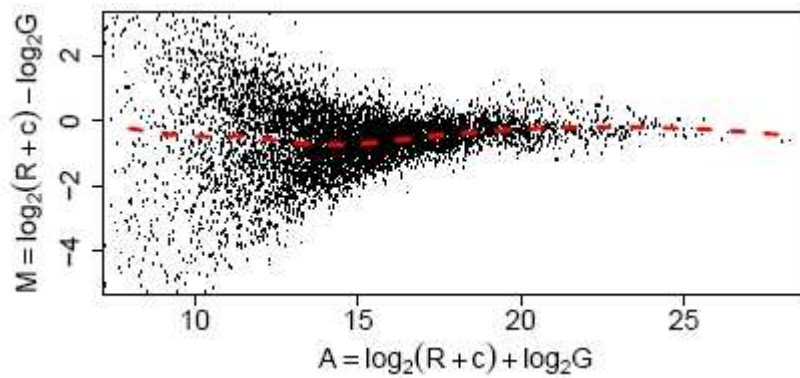


Abbildung 3.2: Ein Scatterplot mit einer doppel-logarithmischen Skala, in welchem dieselben Daten wie in Abb.3.1 aufgetragen sind [1].

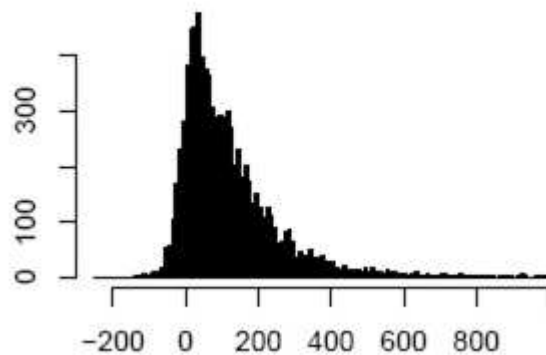


Abbildung 3.3: Ein Histogramm der Intensitäten von quasi-zufällig ausgewählten Genen, dass die unimodale Verteilung zeigt [1]. Y-Achse: Anzahl Gene, X-Achse: gemessene Intensitäten ohne Backgroundnoise.

# 4 Kalibrierung

Bei einem Microarray-Experiment werden die gemessenen Daten in eine Matrix aufgetragen. Dabei werden die Gene (Proben) in die Zeilen eingetragen und die Konditionen (Samples) als Spalten dargestellt. Mit den nun vorhandenen Daten lässt sich noch kein Vergleich von Expressionsdaten bewerkstelligen, weil es Unterschiede in der Aufbereitung der Proben, dem Labeling, der Fluoreszenzintensität und -auslesung gibt. Daher bedarf es einer Kalibrierung (auch Normalisierung genannt). Typischerweise wächst die Varianz der gemessenen Intensitäten an einem Spot mit der Intensität.

## 4.1 Modell nach Chen

Chen (siehe [1], Kapitel 3.1) geht von der Annahme aus, dass die Varianz der Genexpression linear mit dem Erwartungswert steigt. Um sein Modell der Intensitätsdaten von einem 2-farben cDNA Microarray anzupassen, verwendet er eine multiplikative Kalibrierung, wobei der Koeffizient der Abweichung mit einem iterativen Algorithmus abgeschätzt wird. Das Modell von Chen motiviert zur Benutzung von logarithmierten Intensitäten. Dieses Konzept weist folgende Schwächen auf, die aus dem Graphen der Logarithmusfunktion ersichtlich werden, siehe Abb. 4.1:

- Nach einer Nutzung dieses Modells bei vielen Datensätzen ergab sich, dass die Aussagekraft einer bestimmten Rate, die innerhalb einer niedrigen Intensität beobachtet wird, höher ist als innerhalb der hohen Intensität.
- Die Logarithmierung erlaubt keine Transformation von nicht-positiven Werten. Zur Behebung dieses Problems setzt man eine Schwelle für minimale Werte ein, sodass Werte unterhalb der Schwelle nicht beachtet werden. Jedoch gibt es keinen plausiblen Schwellenwert und keine gute Möglichkeit die ausgelassenen Werte in Betracht zu ziehen.

Das Problem dieses Ansatzes liegt in der Tatsache, dass das Varianz-Mittelwert Verhältnis anders ist als von Chen angenommen.

Um dieses Problem zu bewältigen wird nun ein neues Modell von Huber et al. [2] vorgestellt. Die Hauptkomponente dieses Modells ist das von Rocke und Durbin [5] vorge-

schlagene Fehlermodell, die zu einer quadratischen Abhängigkeit der Varianz vom Mittelwert führt. Huber entwickelte eine Familie von Transformationen, so dass die Varianz der transformierten Werte annähernd unabhängig vom Mittelwert ist. Diese Transformationen bilden ein statistisches Modell, welches eine Maximum-Likelihood (ML)-Abschätzung ihrer Parameter erlaubt. Zusätzlich verallgemeinert dieses Modell die Anzahl der Samples  $d$ . Chen ging von  $d = 2$  aus, d.h. vom 2-farben cDNA-Ansatz. Huber ermöglicht durch diese Verallgemeinerung eine Betrachtung einer Serie von einfarbenen ( $d > 2$ ) Microarrayexperimenten.

## 4.2 Modell nach Huber

2001 haben Rocke und Durbin ein Fehlermodell veröffentlicht, in welchem sie annehmen, dass die gemessene Intensität  $y$  eine Realisierung der Zufallsvariable  $Y$ , welche folgendermaßen festgelegt ist:

$$Y = \alpha + \beta e^\eta + \nu, \quad (4.1)$$

$\alpha$  ist dabei der Offset und  $\beta$  die tatsächliche Expression.  $e^\eta$  und  $\nu$  sind jeweils der multiplikative und der additive Fehlerterm, wobei  $\eta$  und  $\nu$  mit  $N(0, 1)$  normal verteilt sind.

Als Erwartungswert von  $Y$  ergibt sich somit:

$$\begin{aligned} E(Y) &= E(\alpha + \beta e^\eta + \nu) \\ &= E(\alpha) + E(\beta e^\eta) + E(\nu) \\ &= \alpha + \beta m_\eta \end{aligned} \quad (4.2)$$

wobei  $m_\eta$  der Erwartungswert von  $e^\eta$  ist.

Daraus folgt für die Varianz von  $Y$ :

$$\begin{aligned} Var(Y) &= E(Y - E(Y))^2 \\ &= E(\alpha + \beta e^\eta + \nu - (\alpha + \beta m_\eta))^2 \\ &= E(\nu + \beta(e^\eta - m_\eta))^2 \\ &= E(\nu^2 + 2\nu\beta(e^\eta - m_\eta) + \beta^2(e^\eta - m_\eta)^2) \\ &= E(\nu^2) + \beta^2 E(e^\eta - m_\eta)^2 \\ &= s_\nu^2 + \beta^2 s_\eta^2 \end{aligned} \quad (4.3)$$

wobei  $s_\nu^2$  und  $s_\eta^2$  jeweils die Varianzen von  $\nu$  und von  $e^\eta$  sind.

Formt man nun die Gleichung 4.2 nach  $\beta$  um, und setzt es in die Gleichung 4.3 ein, so erhält man einen Term der Form

$$v(u) = (c_1 u + c_2)^2 + c_3, \quad \text{mit } c_3 > 0 \quad (4.4)$$

wobei  $v(u)$  die Varianz als Funktion vom Erwartungswert  $u$ ,  $c_1 = \frac{s_\eta}{m_\eta}$ ,  $c_2 = -\frac{\alpha s_\eta}{m_\eta}$  und  $c_3 = s_\nu^2$  sind. Hier wird die quadratische Abhängigkeit der Varianz vom Mittelwert ersichtlich.

Gesucht ist nun eine Transformation  $h(y)$ , für die gilt  $Var(h(y)) = konst.$  Die Methode, welche man anwendet um  $h(y)$  zu finden, ist die sogenannte Delta-Methode. Dabei approximiert man  $h(y)$  um den Erwartungswert  $u$  mithilfe der ersten beiden Glieder der Taylor-Reihe:

$$\begin{aligned} h(y) &\approx h(u) + (y - u)h'(u) \\ &\approx h(u) - uh'(u) + yh'(u) \end{aligned} \quad (4.5)$$

Daraus folgt nun für den Erwartungswert:

$$\begin{aligned} E(h(Y)) &\approx E(h(u) - uh'(u) + Yh'(u)) \\ &= h(u) - uh'(u) + h'(u)E(Y) \\ &= h(u) - uh'(u) + uh'(u) \\ &= h(u) \end{aligned} \quad (4.6)$$

Für die Varianz gilt:

$$\begin{aligned} Var(h(Y)) &= E(h(Y) - E(h(Y)))^2 \\ &\approx E(h(u) - uh'(u) + Yh'(u) - h(u))^2 \\ &= E(Yh'(u) - uh'(u))^2 \\ &= E((h'(u))(Y - u))^2 \\ &= h'(u)^2 \cdot E((Y - u))^2 \\ &= h'(u)^2 \cdot Var(Y) \\ &= h'(u)^2 \cdot v(u) \end{aligned} \quad (4.7)$$

Da man die  $Var(h(Y))$  konstant halten möchte, ergibt sich zusammen mit 4.4:

$$1 = h'(u)^2 \cdot v(u) \quad (4.8)$$

$$h'(u) = \frac{1}{\sqrt{v(u)}} \quad (4.9)$$

$$h(y) = \int^y \frac{1}{\sqrt{v(u)}} du = \int^y \frac{1}{\sqrt{(c_1 u + c_2)^2 + c_3}} du \quad (4.10)$$

Führt man die Integration der Gleichung 4.10 durch, so erhält man:

$$h(y) = \gamma \cdot \operatorname{arsinh}(a + by) \quad (4.11)$$

wobei  $\gamma = c_1^{-1} = \frac{m_\eta}{s_\eta}$ ,  $a = \frac{c_2}{\sqrt{c_3}} = -\frac{\alpha s_\eta}{s_\nu m_\eta}$  und  $b = \frac{c_1}{\sqrt{c_3}} = \frac{s_\eta}{s_\nu m_\eta}$  sind. Der Parameter  $\gamma$  kann weggelassen werden, da es lediglich ein Skalierungsfaktor ist. Näheres zu dieser Varianzstabilisierungstransformation ist in [6] nachzulesen.

Will man nun die Varianzstabilisierungstransformation auf die gemessenen Daten anwenden, muss man beachten, dass die Parameter  $a$  und  $b$  (aus 4.11) für jedes Sample unterschiedlich sind. Somit ergibt sich für jedes Sample  $i$ , für  $i \in \{1, \dots, d\}$  die Varianzstabilisierungstransformation:

$$h_i(y_{ki}) = \operatorname{arsinh}(a_i + b_i y_{ki}) \quad (4.12)$$

wobei die Parameter  $a_i$  und  $b_i$  mithilfe eines Maximum-Likelihood-Abschätzers (s. Kapitel 5) angenähert werden können.

Der Vorteil dieser Transformation gegenüber dem Logarithmus ist das fehlen der Singularität bei 0 (siehe Abb. 4.1).

Zum Vergleich von Gewebe  $i$  zu Gewebe  $j$  der transformierten Genexpressionsraten bei Proben  $k$ , für  $k \in \{1, \dots, n\}$  benutzt man:

$$\Delta h_{k;ij} = \hat{h}_i(y_{ki}) - \hat{h}_j(y_{kj}), \quad \text{für } k = 1, \dots, n \quad (4.13)$$

Für große Intensitäten entspricht dies ungefähr dem Log-ratio, während es für geringe Intensitäten nahe null mehr der Differenz ähnelt.

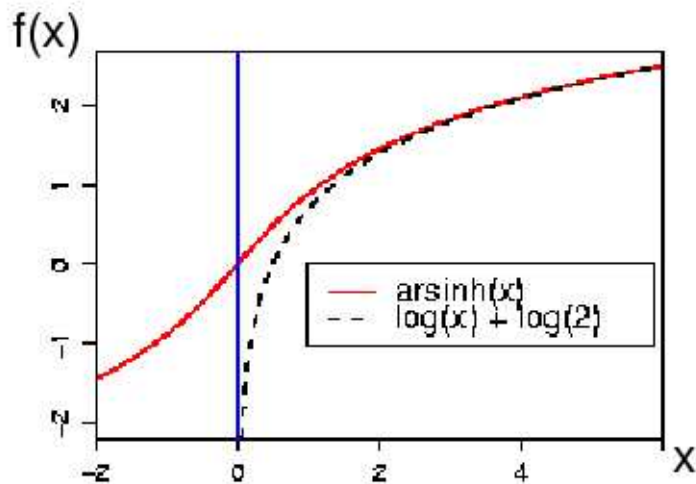


Abbildung 4.1: Graphen der Funktionen  $\log(2x)$  und  $\operatorname{arsinh}(x)$ .

# 5 Parameterabschätzung

Es geht darum, die Parameter  $a_i$  und  $b_i$  der Transformation (4.12) aus den nicht unterschiedlich exprimierten Genen abzuschätzen, welche die geringste Messschwankungen über alle Samples  $i$  aufweisen. Gehen wir davon aus, dass der Fehlerterm  $\varepsilon_{ki}$  normal verteilt ist, erhalten wir

$$h_i(Y_{ki}) = \mu_k + \varepsilon_{ki}. \quad (5.1)$$

wobei  $\mu_k$  der Erwartungswert des Gens  $k$  über alle Samples ist.

Die Parameter werden aus der Menge  $K$  der nicht unterschiedlich exprimierten Gene nach dem Maximum-Likelihood-Prinzip abgeschätzt.  $K$  wird dabei durch einen speziellen Algorithmus von Huber gesucht, welcher auf der LTS (*least trimmed sum of squares*) Regression beruht.

## 5.1 Maximum-Likelihood Abschätzer

Die Likelihood-Funktion, die über die Parameter  $\{a_i\}$ ,  $\{b_i\}$ ,  $c$ ,  $\{\mu_k\}$  maximiert werden soll, ist:

$$\prod_{k=1}^n \prod_{i=1}^d \rho\left(\frac{h_i(y_{ki}) - \mu_k}{c}\right) h'_i(y_{ki}) \quad (5.2)$$

wobei  $\rho(\cdot)$  die Gauß'sche Dichte der Standardnormalverteilung darstellt. Diese Funktion ergibt sich aus dem Produkt über die Wahrscheinlichkeiten der beobachteten Ergebnisse unter der Annahme (5.1). Dabei wird angenommen, dass

$$\begin{aligned} \hat{\mu}_k &= \frac{1}{d} \sum_{i=1}^d h_i(y_{ki}) \\ \hat{c}^2 &= \frac{1}{nd} \sum_{k=1}^n \sum_{i=1}^d (h_i(y_{ki}) - \hat{\mu}_k)^2 \\ \hat{c} &= \hat{c}(a_i, b_i) \end{aligned} \quad (5.3)$$

Setzt man nun (5.3) in (5.2) ein, so erhält man einen profile likelihood-Term, welche man folgendermaßen

$$\begin{aligned}
\text{pl}(a_1, b_1, \dots, a_d, b_d) &= \\
&= \prod_{k=1}^n \prod_{i=1}^d \frac{1}{\sqrt{2\pi\hat{c}}} \exp\left(\frac{(h_i(y_{ki}) - \hat{\mu}_k)^2}{\frac{2}{nd} \sum_{k=1}^n \sum_{i=1}^d (h_i(y_{ki}) - \hat{\mu}_k)^2}\right) h'_i(y_{ki}) \\
&= \frac{e^{nd/2}}{(2\pi)^{nd/2} \hat{c}^{nd}} \prod_{k=1}^n \prod_{i=1}^d h'_i(y_{ki}) \tag{5.4}
\end{aligned}$$

als eine Funktion von den Parameter  $a_i$  und  $b_i$  darstellt. Allerdings wird das Maximum von dem "logarithmierten profile likelihood"

$$\begin{aligned}
\text{pll}(a_1, b_1, \dots, a_d, b_d) &= -nd \log \hat{c} + \sum_{k=1}^n \sum_{i=1}^d \log h'_i(y_{ki}) \\
&= -\frac{nd}{2} \log \left( \sum_{k=1}^n \sum_{i=1}^d (h_i(y_{ki}) - \hat{\mu}_k)^2 \right) \\
&\quad + \sum_{k=1}^n \sum_{i=1}^d \log h'_i(y_{ki}) \tag{5.5}
\end{aligned}$$

gesucht, damit die Produkte in Summen überführt werden können. Das ist zulässig aufgrund der Monotonie von  $\log$ . Die Parameter  $a_i$  und  $b_i$  werden durch eine numerische Maximierung von (5.5) gefunden. Der Maximum-Likelihood Abschätzer ist empfindlich gegenüber der Anwesenheit von unterschiedlich exprimierten Genen.

Nun wird per LTS Regression eine Modifikation durchgeführt, die die Maximum-Likelihood Abschätzung robust gegen Ausreißer macht.

## 5.2 LTS Regression

Die LTS Regression ist eine lineare Regression, die die Summe der Quadrate der Abstände über alle nicht unterschiedlich exprimierten Gene minimiert. Vorteil gegenüber der einfachen linearen Regression (LS Regression) ist, dass die Fehlerrate durch Ausreißer nicht verfälscht wird, solange die Anzahl der Ausreißer auf maximal  $1 - q_{lts}$  beschränkt ist. Dabei ist  $q_{lts}$  der erwartete Anteil der nicht unterschiedlich exprimierten Gene mit relativer Häufigkeit von 0.5 bis 1.

Um das LTS mit dem Maximum-Likelihood Abschätzer zu verbinden, ersetzt man nun die Summen über alle  $k = 1, \dots, n$  von (5.5) durch Summen über die  $k \in K$ . Die Menge der nicht unterschiedlich exprimierten Gene  $K$  wird durch eine iterative Prozedur gewonnen:

1. Zuerst werden die Parameter über alle Gene abgeschätzt, wonach man die Expressionen transformiert.

2. Jetzt werden die Gene nach ihren Erwartungswerten sortiert und in 10 Quantile aufgeteilt.
3. Nun berechnet man für alle Gene eines Quantils den quadratischen Fehler und sortiert die Gene nach der Fehlergröße.
4. Für weitere Berechnungen verwendet man von jedem Quantil den ersten  $q_{lts}$ -Anteil.
5. Die nächste Iteration verwendet Parameter, die man durch eine Abschätzung über die zuletzt gewonnenen Gene gewinnt.

Die Prozedur wird solange iteriert bis eine Stabilisierung der Parameter erfolgt ist. Eine robuste Abschätzung ist dadurch gewährleistet, indem man die gleiche Anzahl von Genen aus jedem Quantil nimmt. In der Praxis reichten 10 Iterationen, um die Parameter erfolgreich abzuschätzen.

# 6 Mustererkennung

Nachdem die Genexpressionsdaten kalibriert und transformiert sind, geht es nun darum, den eigentlichen Ansatz der Analyse dieser Daten weiterzuführen.

Bereits eine Visualisierung der Daten kann erste Aufschlüsse über den Ausgang des Experiments vermitteln. Im folgenden werden weitere Methoden vorgestellt, die einen detaillierteren Einblick verschaffen können.

## 6.1 Projektionsmethoden

Eine wichtige Methodenklasse arbeitet mit Dimensionsreduktion. Die Vektoren der Expressionsmatrix werden dabei in Räume mit weniger Dimensionen projiziert, wobei der Informationsverlust so klein wie möglich gehalten wird. Die Visualisierung der projizierten Daten kann dem Experimentator wichtige Informationen über die Daten geben.

Eine sehr verbreitete Methode der Dimensionsreduktion ist die Hauptkomponentenanalyse (*PCA- principal component analysis*). Dabei werden die Vektoren auf die Eigenvektoren der Datenmatrix mit den größten Eigenwerten projiziert. Die Projektion auf die ersten beiden Hauptkomponenten ist dabei besonders hilfreich, da sie die Möglichkeit des Auftragens in einen 2D-Plot erlaubt. Eine interessante Entdeckung machten dabei Alter et al. Sie zeigten, dass der Eigenvektor mit dem größten Eigenwert die experimentellen Artefakte wiedergibt und somit seine Anteile herausgefiltert werden können.

## 6.2 Clusteralgorithmen

Möchte man aus Expressionsdaten auf funktionelle Verwandtschaft oder Coregulation der Gene schließen, bedient man sich Clustermethoden. Hierbei geht man von der Annahme aus, dass solche Gene ähnliche Expressionsprofile aufweisen. Clusteralgorithmen haben die Gruppierung von Objekten anhand von Ähnlichkeitsmerkmalen als Ziel.

Beim hierarchischen Clustern wird ein Dendrogramm berechnet, welche die Objekte als Blätter enthält. Man unterscheidet dabei zwischen zwei Varianten:

**Bottom-Up-Methoden**, auch agglomerative Methoden genannt, geht von den Blättern aus, und fasst ähnliche Blätter zu Clustern zusammen.

**Top-Down-Methoden**, auch divisive Methoden genannt, betrachtet alle Objekte als Elemente eines großen Clusters, bevor dieser iterativ zerkleinert wird.

Die verschiedenen Cluster sind als Endprodukt beider Varianten im Dendrogramm ersichtlich. Hierarchisches Clustern wird generell dazu verwendet die Zeilen und Spalten der Genexpressionsmatrix in eine Ordnung zu bringen, sodass ähnlich Zeilen und Spalten nah beieinander liegen. Nicht-hierarchische Clusteralgorithmen unterteilen den Datensatz direkt in eine vorgegebene Anzahl von Clustern. Bekannte Beispiele sind das  $k$ -Means Clustering und Self-Organizing Maps. Dabei geht man immer davon aus, dass die Genexpressionsdaten in erkennbaren Gruppen organisiert sind. Nichtsdestotrotz werden die Daten auch dann geclustert, wenn keine Partitionierung in den Daten vorliegt. Daher ist es wichtig, die Anzahl der Cluster im Datensatz möglichst genau abzuschätzen, wofür verschiedene Autoren Methoden entwickelten, um die Validität des Clusterings zu bewerten.

Das Hauptproblem der Clusteralgorithmen ist, dass solche Methoden auf der Basis einer globalen Ähnlichkeitsmessung arbeiten. Jedoch gibt es bestimmte biologische Situationen in denen Ähnlichkeiten in Expressionsmustern von einigen wenigen Gengruppen auftreten. Das Auffinden von solchen lokalen Mustern, welche Gene betreffen, die in einem gemeinsamen Stoffwechselweg arbeiten oder coreguliert werden, wurde von verschiedenen Autoren beschrieben, die desöfteren Score-basierte Methoden verwenden.

# 7 Ergebnisse

Die vorgestellte Transformation ist eine erfolgreiche und mittlerweile anerkannte Methode, die Genexpressionsdaten zu kalibrieren. Das Ergebnis ist aus der Abbildung 7.1 ersichtlich. Diese Visualisierung zeigt in beeindruckender Weise die Güte der Transformation. Die Varianz ist annähernd konstant.

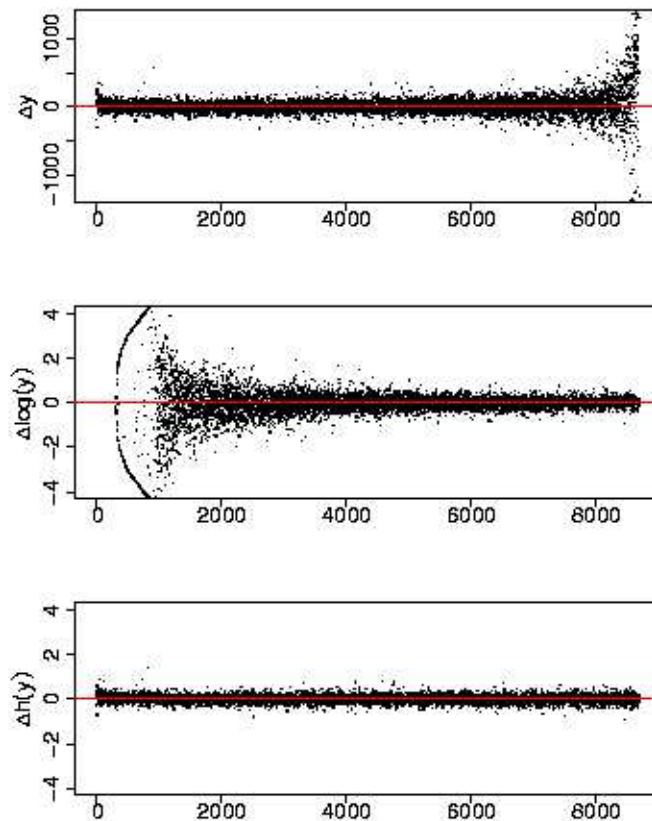


Abbildung 7.1: Drei unterschiedliche Transformationen angewendet auf Daten eines cDNA-Microarray. Die X-Achsen der drei Plots spiegeln jeweils das Intensitätenspektrum wieder. **Oben:** Differenz von Grün und Rot aufgetragen gegen die Summe von Grün und Rot. **Mitte:** Log-ratios. **Unten:** nach Huber transformierte Differenz des grünen und roten Kanals.

## 8 Literaturverzeichnis

- [1] Huber, W., Heydebreck, A. v., Vingron, M. *Analysis of Microarray Gene Expression Data*. Handbook of Statistical Genetics, 2nd edition, Wiley, 2003.
- [2] Huber, W., Heydebreck, A. v., Sültmann, H., Poustka, A., Vingron, M. *Variance Stabilization Applied to Microarray Data Calibration and to the Quantification of Differential Expression*. Bioinformatics, 18 Suppl 1, S96-S104, 2002.
- [3] Huber, W., Heydebreck, A. v., Sültmann, H., Poustka, A. and Vingron, M. *Parameter Estimation for the Calibration and Variance Stabilization of Microarray Data*. Statistical Applications in Genetics and Molecular Biology, Vol. 2, No. 1, Article 3, 2003.
- [4] Rousseuw, P. J., Leroy, A. M. *Robust Regression and Outlier Detection*. John Wiley and Sons, 1987
- [5] Rocke, D. M., Durbin, B. P. *A Model for Measurement Error for Gene Expression Analysis*. Journal of Computational Biology, 8:557-569, 2001
- [6] Tibshirani, R., *Estimating Transformations for Regression via Additivity and Variance Stabilization*. J. Amer. Stat. Assoc., 83:394-405, 1988.
- [7] Durbin, B. P., Hardin, J. S., Hawkins, D. M. Rocke, D. M. *A Variance-Stabilizing Transformation for Gene-Expression Microarray Data*. Bioinformatics, 18 Suppl 1, S105-S110, 2002.
- [8] Quinn, K. *A Review of Some Basic Mathematical Concepts and Differential Calculus*. University of Washington, 2002.