Wie der P-Wert in Microarrayexperimenten berechnet wird

MICHAEL NUHN
Nano+Bio-Center
Technische Universität Kaiserslautern
67653 Kaiserslautern, Germany
nuhn@rhrk.uni-kl.de

9. September 2005

Zusammenfassung

Im folgenden wird gezeigt, wie der geheimnisvolle P-Wert berechnet wird.

1 Das Problem

Ein Microarrayexperiment wird *n* Mal wiederholt. Bei jedem Experiment wird das Ratio aus der Intensität von Cy3 und Cy5 gebildet. So erhält man für jedes Gen im Idealfall *n* Ratios. In der Praxis fallen bei vielen Genen aus verschiedenen technischen Gründen einzelne Werte weg, man erhält also eigentlich nur höchstens *n* Werte.

Falls ein Gen aus den beiden Proben gleich stark exprimiert wird, so müssen die Ratios für dieses Gen alle gleich 1 sein. Aus verschiedenen Gründen wird das praktisch nie der Fall sein. Daher wird man versuchen, aus der Meßreihe den Erwartungswert für die Ratios zu schätzen. Je mehr Meßwerte zur Verfügung stehen, desto genauer wird die Schätzung des Erwartungwertes ausfallen. Diese Schätzung wird auch bei gleich exprimierten Genen praktisch nie genau gleich 1 sein. Bei differentiell exprimierten Genen wird der geschätzte Erwartungswert „weiter von 1 entfernt sein“, bei gleich exprimierten Genen wird der geschätzte Erwartungwert „nahe“ bei 1 sein. Es stellt sich die Frage, ab wann man annehmen sollte, daß ein Gen in den Proben unterschiedlich stark exprimiert wird.

Man muß sich dabei klarmachen, daß eine sichere Aussage nie gemacht werden kann. Auch wenn die Ratios noch so weit von 1 entfernt sind, gibt es dennoch eine sehr kleine aber positive

Wahrscheinlichkeit dafür, daß der Erwartungswert 1 ist und die große Abweichung nur dadurch zustande kam, daß die Stichprobe ein Ausreißer war.

Die Frage die sich daher stellt ist: Wie groß ist die Wahrscheinlichkeit, daß ich diese Werte bei einem Gen bekomme, obwohl es in beiden Proben gleich stark exprimiert wird? Diese Wahrscheinlichkeit nennt man den P-Wert.

Wenn dieser P-Wert sehr klein ist, so wird man annehmen, daß die Gene unterschiedlich stark exprimiert werden. Um den P-Wert zu ermitteln, wird ein Hypothesentest durchgeführt. Es werden für jedes Gen zwei Hypothesen aufgestellt:

H_0 : Der Erwartungswert für die Ratios ist 1
H_1 : Der Erwartungswert für die Ratios ist ungleich 1

Es wird dann die Wahrscheinlichkeit p dafür berechnet, daß der geschätze Erwartungwert zustande kam, obwohl der wahre Erwartungwert 1 ist. Wenn p sehr klein ist, wird man H_0 ablehnen und H_1 annehmen müssen, das Gen also als signifikant differentiell exprimiert ansehen.

2 Die Hölle der Statistik

Im folgenden soll hergeleitet werden, wie der P-Wert berechnet wird. Dazu wird in Abschnitt 2.1 zuerst die Transformation der Daten erklärt. In Abschnitt 2.2 wird beschrieben, wie der Erwartungswert und die Varianz der Stichprobe geschätzt werden kann. Abschnitt 2.3 leitet her, wie zuverlässig diese Schätzung ist und in Abschnitt 2.4 wird dann daraus die Formel zur Berechnung des P-Wertes hergeleitet.

2.1 Transformation der Daten

Zu jedem Gen seien k Ratios ($2 < k \leq n$) gegben:

Genname: $Ratio_1$ $Ratio_2$ ... $Ratio_k$

Zu jedem Gen wird der Logarithmus betrachtet.

$$X_i = log(Ratio_i)$$

Die einzelnen Werte werden als Zufallsvariablen X_i modelliert. Alle X_i folgen derselben Normalverteilung mit unbekanntem Erwartungswert μ und unbekannter Varianz σ^2:

$$X_i \sim N(\mu, \sigma^2)$$

Daß die Logarithmen der Ratios normalverteilt sind, wurde empirisch festgestellt. Von verschiedenen Microarrayexperimenten wurden Histogramme der Daten ausgewertet und es zeigte sich, daß deren Verteilung am besten mit einer Normalverteilung angenähert werden kann.

2.2 Schätzen der Parameter der Verteilungsfunktion

Die Normalverteilung hat zwei Parameter, den Erwartungswert μ und die Varianz σ^2. Diese Werte sind bei einem Microarrayexperiment unbekannt, können aber aus aus den Meßwerten geschätzt

werden. Ein erwartungstreuer Schätzer ist eine Funktion, die aus den Meßwerten eine Schätzung für einen dieser Parameter liefert.

Ein erwartungstreuer Schätzer für den Erwartungswert ist bei normalverteilten Zufallsvariablen das Stichprobenmittel:

$$M := \frac{1}{k}\sum_{i=1}^{k} X_i$$

Es ist naheliegend, die Varianz analog zu ihrer Definition folgendermaßen zu schätzen:

$$\bar{V} = \frac{1}{k}\sum_{i=1}^{k}(X_i - M)^2 \tag{1}$$

Wäre der Erwartungswert μ bekannt und stünde dieser in (1) anstelle der Schätzung M, so wäre dies ein erwartungstreuer Schätzer für die Varianz der X_i. Da der Erwartungswert nicht vorliegt sondern nur das Stichprobenmittel, ist der Schätzer (1) nicht erwartungstreu. Es muß stattdessen die Stichprobenvarianz verwendet werden:

$$\bar{V} := \frac{1}{k-1}\sum_{i=1}^{k-1}(X_i - M)^2$$

2.3 Der geschätzte Erwartungswert und seine Verteilung

Der Mittelwert M einer Meßreihe ist eine Zufallsvariable, die von den X_i abhängig ist. Beim Hypothesentest später soll der Frage nachgegangen werden, mit welcher Wahrscheinlichkeit die Schätzung um einen bestimmten Betrag vom wahren Erwartungswert differieren kann. Um dies berechnen zu können, muß die Verteilung von M ermittelt werden. Da die X_i normalverteilt sind, ist M ebenfalls normalverteilt. Der Erwartungwert $E(M)$ ist wegen

$$\begin{aligned} E(M) &= E(\tfrac{1}{k}\textstyle\sum_{i=1}^{k} X_i) \\ &= \tfrac{1}{k} E(\textstyle\sum_{i=1}^{k} X_i) \\ &= \tfrac{1}{k}\textstyle\sum_{i=1}^{k} E(X_i) \\ &= \tfrac{1}{k}\textstyle\sum_{i=1}^{k} \mu \\ &= \tfrac{1}{k} k\mu = \mu \end{aligned}$$

ebenfalls μ. Um die Varianz von M herleiten zu können, seien hier ohne Beweis zwei Rechenregeln für die Varianz angegeben. A und B seien dabei zwei stochastisch unabhängige Zufallsvariablen und $k \in \Re$ ein Skalar. Dann gilt:

$$V(kA) = k^2 V(A)$$

und

$$V(A+B) = V(A) + V(B).$$

Die Varianz von M ist daher:

$$\begin{aligned} V(M) = &\quad V(\tfrac{1}{k}\textstyle\sum_{i=1}^{k} X_i) \\ = &\quad (\tfrac{1}{k})^2 V(\textstyle\sum_{i=1}^{k} X_i) \\ = &\quad (\tfrac{1}{k})^2 \textstyle\sum_{i=1}^{k} V(X_i) \\ = &\quad (\tfrac{1}{k})^2 \textstyle\sum_{i=1}^{k} \sigma^2 \end{aligned}$$

$$= (\frac{1}{k})^2 k\sigma^2 = \frac{\sigma^2}{k} \tag{2}$$

Die Varianz von M sinkt also linear mit der Anzahl der vorliegenden Meßwerte. Sie wird umso geringer sein, je mehr Messungen gemacht werden.

2.4 Der Hypothesentest

Die Hypothesen H_0 und H_1 lauten für die logarithmierten Ratios:

H_0 : Der Erwartungswert für die logarithmierten Ratios ist 0.
H_1 : Der Erwartungswert für die logarithmierten Ratios ist ungleich 0.

Es soll ermittelt werden, wie wahrscheinlich es ist, diese Schätzung des Erwartungswertes zu erhalten, wenn H_0 wahr ist. Die Wahrscheinlichkeit $P(-c \leq M \leq c)$, daß M im Intervall $[-c..c]$ liegt, ist die Fläche unter der Dichtefunktion im Intervall $[-c..c]$ in Abbildung 1.

Die Wahrscheinlichkeit, daß M weiter als c von 0 entfernt ist, also außerhalb des Intervalls $[-c..c]$ liegt, ist unter der Annahme von H_0:

$$P(|M-0| \geq c) = 1 - P(-c \leq M \leq c) \tag{3}$$

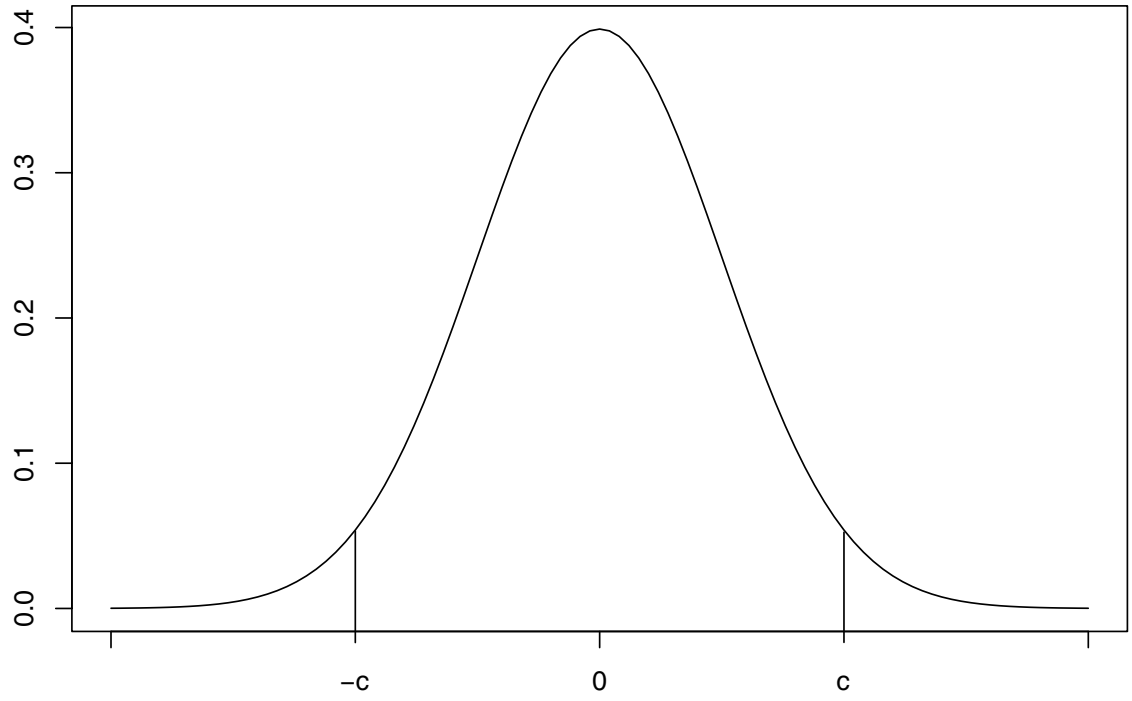
M nehme nun bei einer Meßreihe den Wert M_0 an. Um jetzt die Wahrscheinlichkeit zu erhalten, daß M unter der Bedingung H_0 mindestens so weit von 0 entfernt ist wie es M_0 ist, setzt man $c := |M_0 - 0|$ und erhält mit (3):

$$\begin{aligned} P(|M-0| \geq |M_0-0| \quad |H_0) = &\quad 1 - P(0 - |M_0-0| \leq M \leq 0 + |M_0-0| \qquad |H_0) \\ = &\quad 1 - P(-|M_0-0| \leq M-0 \leq |M_0-0| \qquad |H_0) \end{aligned}$$

$$= 1 - P(-\frac{|M_0-0|}{\sqrt{\frac{\bar{V}}{n}}} \leq \frac{M-0}{\sqrt{\frac{\bar{V}}{n}}} \leq \frac{|M_0-0|}{\sqrt{\frac{\bar{V}}{n}}} \quad |H_0) \tag{4}$$

$\bar{V}$ ist dabei eine Schätzung für die Varianz σ^2 der X_i. Der Ausdruck $\sqrt{\frac{\bar{V}}{n}}$ ist nach Gleichung (2) eine Schätzung für die Standardabweichung von M. Da die Anzahl der Meßwerte klein ($k < 30$) ist, ist die Schätzung $\sqrt{V}$ von σ nicht gut genug, um die Verteilung von $\frac{M-0}{\sqrt{\frac{\bar{V}}{n}}}$ in Gleichung (4) mit der Standardnormalverteilung $N(0,1)$ anzunähern.

Abbildung 1: Dichte von M unter der Bedingung H_0



Die Verteilung von $\frac{M-\mu}{\sqrt{\frac{\bar{V}}{n}}}$ ist die Studentverteilung mit $k-1$ Freiheitsgraden:

$$\frac{M-\mu}{\sqrt{\frac{\bar{V}}{k}}} \sim Student_{k-1}$$

Unter Annahme von H_0 ist $\mu = 0$ und Gleichung (4) kann berechnet werden:

$$\begin{aligned} P(|M-0| \geq |M_0-0| \quad |H_0) \quad &= 1 - \int_{-\frac{|M_0-0|}{\sqrt{\frac{\bar{V}}{n}}}}^{\frac{|M_0-0|}{\sqrt{\frac{\bar{V}}{n}}}} df(x)dx \\ &= 1 - (2pt(\frac{|M_0-0|}{\sqrt{\frac{\bar{V}}{n}}}) - 1) \\ &= 2(1 - pt(\frac{|M_0-0|}{\sqrt{V}}\sqrt{n}) \end{aligned} \tag{5}$$

pt ist dabei die Verteilungsfunktion der Studentverteilung mit $k-1$ Freiheitsgraden.

Gleichung (5) ist die Wahrscheinlichkeit dafür, daß der geschätzte Wert M_0 erscheint, obwohl H_0 wahr ist. Das ist der P-Wert. Ist dieser sehr gering, so ist die Wahrscheinlichkeit dafür, daß die Meßreihe unter der Bedingung H_0 zustande kam, sehr gering. In dem Fall würde man die

Hypothese H_0 ablehnen und H_1 annehmen, die besagt, daß der Erwartungswert der logarithmierten Ratios nicht 0 ist. Ein solches Gen würde man als signifikant differentiell exprimiert bezeichnen.

Bei Microarrayexperimenten muß man beachten, daß es sich um eine sehr große Zahl von Einzelexperimenten handelt. Auch wenn man H_0 nur für Gene ablehnt, die einen P-Wert von von $< 1\%$ haben, so würde man bei 2000 Genen, deren Ratios einen Erwartungswert von 1 haben, die Hypothese im Durchschnitt H_0 $2000 * 0.01 = 20$ Mal ablehnen, obwohl sie wahr ist, also 20 Gene fälschlicherweise als signifikant betrachten.

Dieses Problem kann man durch angepaßte P-Werte angehen. Die angepaßten P-Werte sind genaugenommen keine mehr sondern schätzen die False discovery rate oder die Family wise error rate. Diese angepaßten P-Werte werden durch eine Analyse der P-Werte berechnet und sollten bei der Auswertung von Microarraydaten mit hinzugezogen werden.