

Comparison of the Empirical Bayes and the Significance Analysis of Microarrays ^[1]

Jens Keienburg
Molekulare Biotechnologie / Bioinformatik
Universität Heidelberg
Kontakt : JensKeienburg@gmx.de

02.05.2005

Comparison of EBAM and SAM

Übersicht

- Statistische Grundlagen
 - t-test
 - Wilcoxon Test
- Analyse von Microarrays
 - SAM (t-Statistik)
 - SAM (Wilcoxon-Statistik)
 - EBAM (t-Statistik)
 - EBAM (Wilcoxon-Statistik)
- Vergleich
- Zusammenfassung und Wiederholung
- (Fragen, Diskussion und Übungen je nach Bedarf)

t-Test

- Definitionen :

$X := \text{Zufallsvariable}$

$H := \text{Hypothesenraum}$

$K := \text{Alternative}$

$\Theta := H \cup K$

- Test : $\Phi(x) : \chi \rightarrow [0, 1]$, wobei χ der Ereignisraum von X ist.

$\Phi(x) = 1 \Rightarrow$: Hypothese wird abgelehnt

$\Phi(x) = 0 \Rightarrow$: Hypothese wird nicht abgelehnt

- Likelihood-Quotienten : $q(x) = \frac{\sup\{L_x(\theta) : \theta \in K\}}{\sup\{L_x(\theta) : \theta \in H\}}$

- Gesucht ist ein $c \in \mathbb{R}$, so dass :

$q(x) > c \Leftrightarrow \Phi(x)$ ($q(x) > c$ ist hier ein bool'scher Ausdruck)

t-Test

- X_1, \dots, X_n seien $N(\mu, \sigma^2)$ -verteilte Zufallsvariablen mit unbekanntem (μ, σ^2) . Für gegebenes μ_0 ist zu testen, ob $\mu_0 = \mu$ oder $\mu_0 \neq \mu$ ist. Es gilt:

$$\sup\{L_x(\theta) : \theta \in K\} = f(x | (\bar{x}, \hat{\sigma}^2))$$

$$\sup\{L_x(\theta) : \theta \in H\} = f(x | (\mu_0, \tilde{\sigma}^2))$$

$$f(x | (\bar{x}, \hat{\sigma}^2)) = \frac{1}{\hat{\sigma}\sqrt{2\pi}} e^{-(x-\bar{x})^2/\hat{\sigma}^2}$$

$$f(x | (\bar{x}, \tilde{\sigma}^2)) = \frac{1}{\tilde{\sigma}\sqrt{2\pi}} e^{-(x-\bar{x})^2/\tilde{\sigma}^2}$$

$$\bar{x} = \frac{1}{n} \sum_i x_i$$

$$\hat{\sigma}^2 = \frac{1}{n} \sum_i x_i - \bar{x}$$

$$\tilde{\sigma}^2 = \frac{1}{n} \sum_i x_i - \mu_0$$

t-Test

- $q(x) = \frac{\sup\{Lx(\theta):\theta\in K\}}{\sup\{Lx(\theta):\theta\in H\}} = \frac{f(x|(\bar{x},\hat{\sigma}^2))}{f(x|(\mu_0,\tilde{\sigma}^2))} = \dots = 1 + \frac{(\bar{x}-\mu_0)^2}{\hat{\sigma}^2}$

Für geeignetes c ist $q(x) \geq c \Leftrightarrow \phi(x)$

- Definition von $T(x)$:

$$s(x) = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$$

$$T(x) := \frac{\sqrt{n}(\bar{x}-\mu_0)}{s(x)}$$

Für geeignetes $t \in \mathbb{R}$ ist $T(x) \geq t \Leftrightarrow q(x) \geq c \Leftrightarrow \phi(x)$

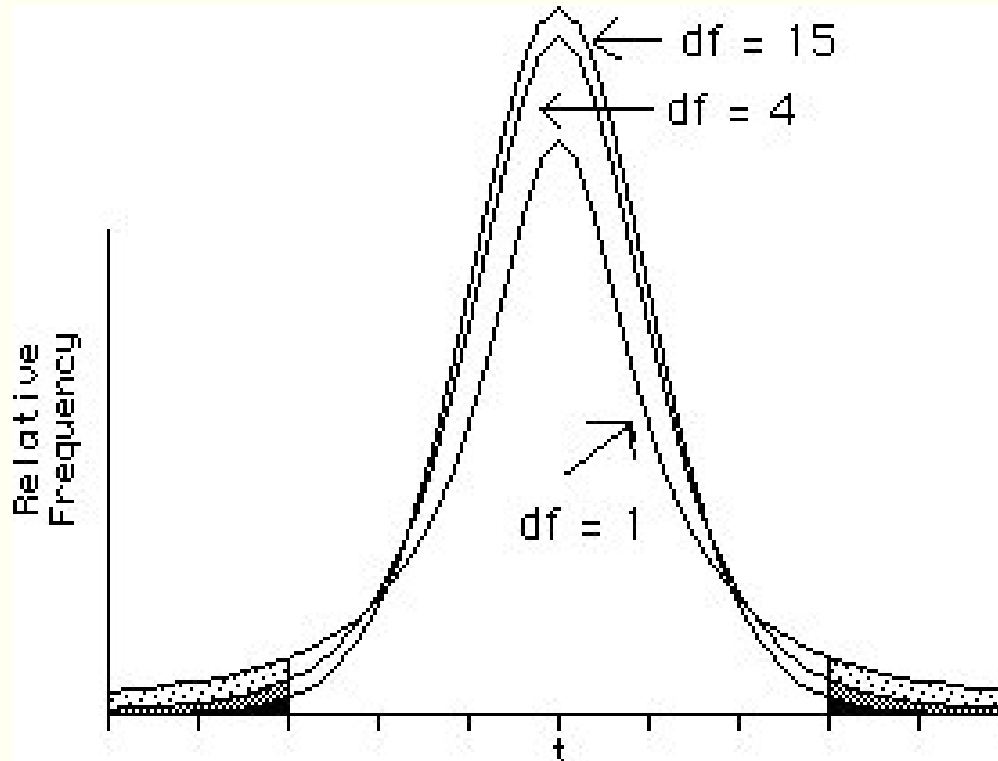
t-Test

- Seien $Y_i = (X_i - \mu_0)/\sigma$ normalverteilte und nur von X_i abhaengige Zufallsvariablen, dann gilt :

$$T(X) = \frac{\bar{Y}\sqrt{n}}{s(Y)}$$

- $s(y)$ ist $\chi^2/(n - 1)^2$ -verteilt. Hieraus lässt sich die Verteilung von $T(x)$ ableiten. Es lässt sich zeigen, dass diese nur von n abhängt.
- Satz : Sind X_1, \dots, X_n unabhängige $N(\mu_0, \sigma^2)$ -verteilte Zufallsvariablen und ist $T(X)$ wie oben definiert, dann ist $T(X)$ t_{n-1} -verteilt.

t-Test



t-Test

- Seien $X = X_1, \dots, X_m$ $N(\mu_1, \sigma_1^2)$ - und $Y = Y_1, \dots, Y_n$ $N(\mu_2, \sigma_2^2)$ -verteilt. Die Nullhypothese $\theta \in H$ sei, dass $\mu_1 = \mu_2$, und die Alternative $\theta \in K$, dass $\mu_1 \neq \mu_2$ ist.
- Unter der Hypothese ist

$$T(X) := \frac{\bar{X} - \bar{Y}}{s \sqrt{\frac{1}{m} + \frac{1}{n}}}$$

$$\text{mit } s^2 := s^2(X, Y) = \frac{1}{m+n-2} \left(\sum_{i=1}^m (X_i - \bar{X})^2 + \sum_{i=1}^n (Y_i - \bar{Y})^2 \right)$$

t_{n+m-2} -verteilt.

- Bei gegebenem Signifikanzniveau α gibt es β -Quantil $t_{k,\beta}$ ($\beta = 1 - \alpha$), so dass

$$|T(X, Y)| > t_{m+n-2,\beta}$$

die Hypothese mit Signifikanzniveau α verworfen wird.

t-Test

- Wiederholung
- gegebenenfalls Aufgabe :

Gemessen werden 15 Genexpressionswerte in einem zellphysiologisch identischen Zustand. Der Mittelwert betraegt \bar{x} und die geschätzte Varianz s^2 . Wie lautet das Konfidenzintervall fuer den Erwartungswert μ zum Niveau $\alpha = 0.1$

t-Test

- Lösung : Die Genexpressionswerte sind t_{14} -verteilt. Die zugehörige T-Funktion lautet :

$$T(x) = \frac{\sqrt{15}(\bar{x} - \mu)}{s(x)}$$

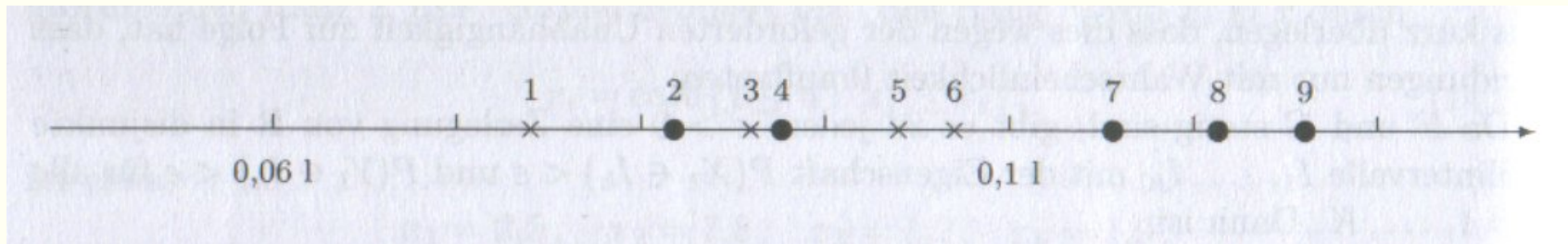
Die beiden zum Signifikanzniveau α zugehörigen β -Quantile sind $t_{14, 1 - \frac{\alpha}{2}}$ und $t_{14, \frac{\alpha}{2}}$. Das Konfidenzintervall lautet damit :

$$[T^{-1}(t_{14, 0.95}); T^{-1}(t_{14, 0.05})]$$

Wilcoxon-Test

- Für einen gegebenen Datenvektor $X = x_1, \dots, x_n$ bezeichne $x_{(1)}, \dots, x_{(n)}$ die zugehörige Ordnungsstatistik und r_i den Rang von x_i .
- Seien X_1, \dots, X_m und Y_1, \dots, Y_n Zufallsvariablen mit unbekanntem Wahrscheinlichkeitsverteilungen f_X und f_Y .
Hypothese: $f_X = f_Y$, bzw. $f_X \neq f_Y$ mit $P(Y_j > t) \leq P(X_i > t), \forall t$
Alternative: $f_X \neq f_Y$ mit $P(Y_j > t) \geq P(X_i > t), \forall t$
- Unter der Hypothese gilt, dass für eine beliebige Permutation π der Zufallsvariablen $Z := (X_1, \dots, X_m, Y_1, \dots, Y_n)$, die Wahrscheinlichkeit des Ereignisses $A_\pi = \{\omega \in \Omega : Z_{\pi(1)}(\omega) < \dots < Z_{\pi(n)}(\omega)\}$ gerade $P(A_\pi) = \frac{1}{N!}$

Wilcoxon-Test



- Im Falle der Alternative sind Y_j tendenziell grösser als X_i . Die Rangsumme $W = \sum_i r_i$, für $i = 1, \dots, m$ von X ist dann entsprechend kleiner.
- Zu einem gegebenen Signifikanzniveau α gibt es ein $c(\alpha, m, n)$, so dass $P(W \leq c) \leq \alpha$ ist. Dieser Wert lässt sich in Tabellen nachschlagen.
- $(W(x) \leq c) \Leftrightarrow \Phi(x)$

Microarray Analyse

- Definitionen : Gegeben ist eine Matrix X_{ij} , welche $j = 1, \dots, J$ Genexpressionswerte fuer $i = 1, \dots, I$ Gene gespeichert hat. n_1 Expressionswerte wurden unter Bedingung 1 gemessen ($j = 1, \dots, n_1$) und n_2 unter Bedingung 2 ($j = 1 + n_1, \dots, J$).

d_i bezeichnet den Scorewert, der entsprechend der T-Statistik berechnet wird :

$$d_i = \frac{\bar{X} - \bar{Y}}{s_i + s_0}$$

s_i ist die Standardabweichung und s_0 ist der sogenannte "fudge"-Faktor.

FDR

- allgemein : $FDR = E\left(\frac{V}{R}\right)$
- Signifikanzniveau $\alpha = 0.05$ und Proben=3000 $\Rightarrow FDR = 150$
- $\widehat{FDR} = \frac{\hat{\pi}_0 \alpha m}{\#\{p_i | p_i \leq \alpha\}}$

p_i : p-value of gene i.

π_0 : a prior Wahrscheinlichkeit für die Hypothese

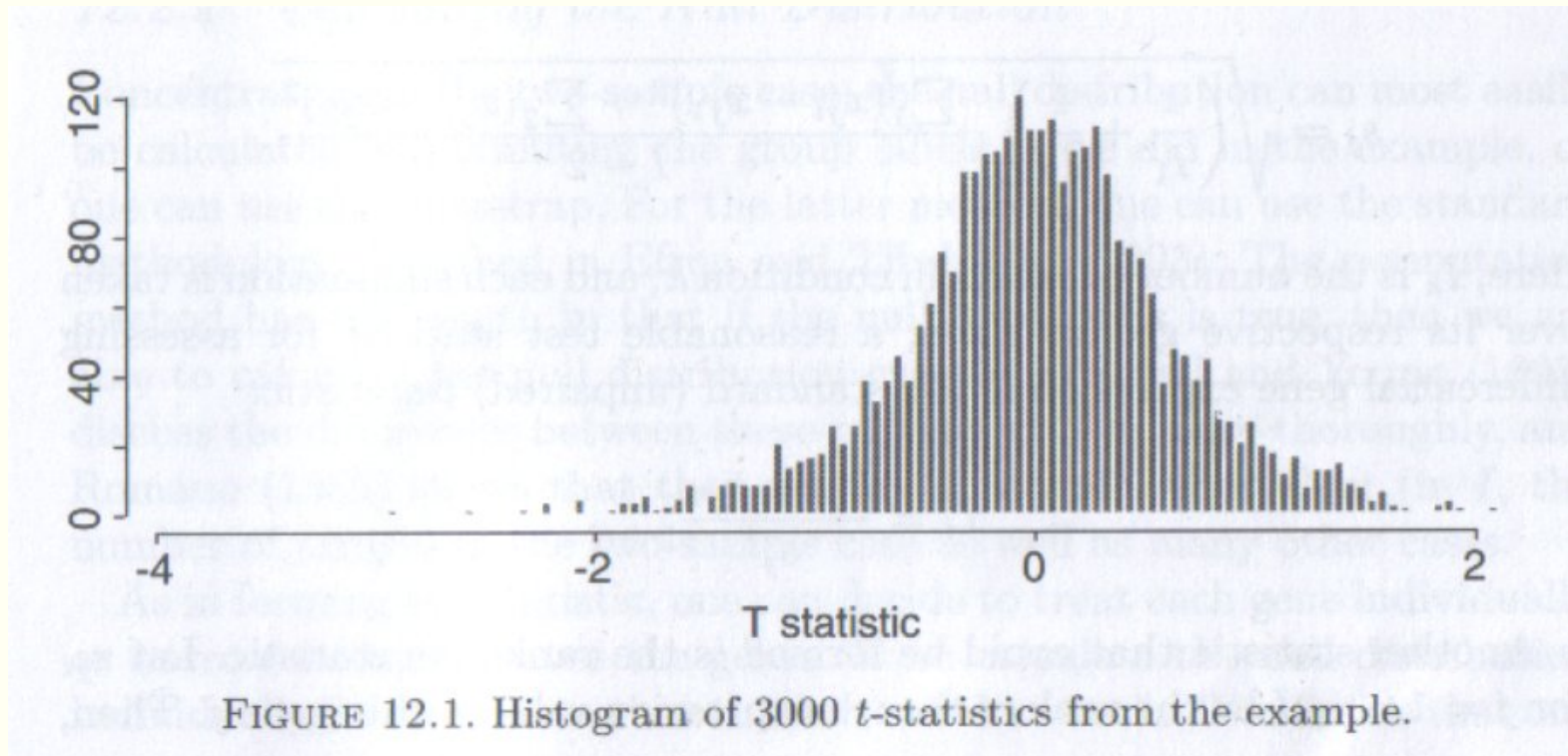
- Berechnung von $\hat{\pi}_0$: Betrachtung der Anzahl der p-Values nahe bei 1.
 $\hat{\pi}_0 = \frac{\#\{p_i | p_i > \alpha\}}{(1-\alpha) \cdot m}$, fuer $\alpha \approx 1$

Gegebenenfalls verbesserte Schätzung durch Extrapolation bei $\alpha = 1$.

FDR - Rechenbeispiel

- Gegeben sind :
 $J = 3000$ Gene
Gruppe 1 (normal) 15 Proben
Gruppe 2 (behandelt) 13 Proben, für alle Gene $j = 1, \dots, J$
Identität der Proben für alle Gene : $(n, n, \dots, n, b, b, \dots, b)$
Label der Proben für alle Gene : $(1, 1, \dots, 1, 2, 2, \dots, 2)$
- Bestimmung der T-Statistik und des β -Quantils $t_{k,\beta}$ für gegebenes β
$$T(j) = d_j = \frac{\bar{x}_{j2} - \bar{x}_{j1}}{s_j}, \text{ für } j = 1, \dots, 3000$$

Histogramm der T-Statistik $T(d_j)$



Berechnung der FDR

- Definitionen :

$$t_{k,\beta} = 2$$

V := Anzahl der falsch positiven Hypothese

S := Anzahl der richtig positiven Hypothesen

$R := V + S =$ Anzahl der signifikanten Hypothesen d_j

- $\widehat{R} = |\{d_j | d_j \geq t_{k,\beta}\}| = 146$

- Wie groß ist die FDR innerhalb dieser 146 Proben ?

Schätzung durch Berechnung der Scores d_j^b für $b = 1, \dots, 100$ bei einer zufälligen Permutationen der Probenlabel $(1, 1, \dots, 1, 2, 2, \dots, 2)$.

Schätzer von π_0 und FDR

- π_0 bezeichnet die a priori Wahrscheinlichkeit für die Nullhypothese.
- Für alle $i = 1, \dots, I$ und $b = 1, \dots, B$ gilt

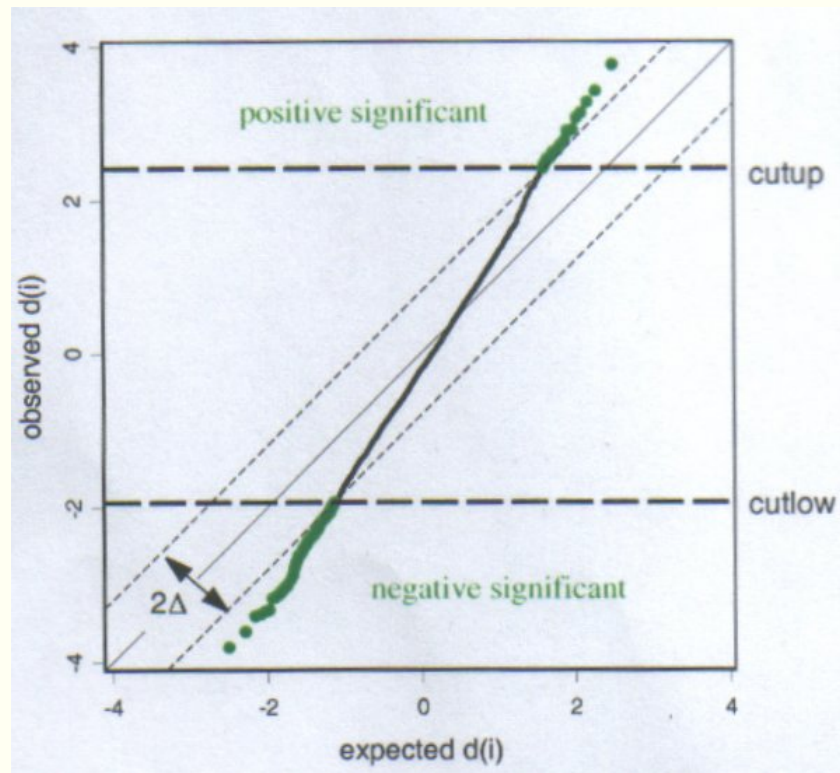
$$\hat{\pi}_0 = \frac{|\{d_i | d_i < 0.15\}|}{|\{d_i^b | d_i^b < 0.15\}|/B} = 0.89$$

$$\widehat{FDR} = \hat{\pi}_0 \frac{|\{d_i^b | d_i^b \geq t_{k,\beta}\}|/B}{|\{d_i | d_i \geq t_{k,\beta}\}|} = 0.89 \cdot 8.42\% = 7.49\%$$

Statistical Analysis of Microarrays (SAM)

- Software, die basierend auf einer modifizierten t-Statistik statistisch signifikante Geneexpressionen und eine entsprechende FDR ermittelt.
- Algorithmus :
 1. Berechnung des Expression-Scores d_i für alle Gene und Erstellung der Rangordnung $d_{(1)} \leq \dots \leq d_{(m)}$.
 2. Berechnung der Score d_i^b für B Permutationen der Probenlabel, Berechnung der Rangordnung $\hat{d}_{(1)}^b \leq \dots \leq \hat{d}_{(m)}^b$ mit $\hat{d}_{(i)}^b = \sum_b d_{(i)}^b / B$
 3. Auftragung von $d_{(i)}$ gegen $\hat{d}_{(i)}^b$ (SAM Plot)
 4. Bestimmung von $cut_{up}(\Delta)$ und $cut_{down}(\Delta)$ bei gegebenem Schwellenwert $\Delta > 0$. Alle Gene mit $d_i > cut_{up}(\Delta)$ bzw. $d_i < cut_{down}(\Delta)$ sind positiv signifikant.

Statistical Analysis of Microarrays (SAM)



SAM - Bestimmung der FDR

- Schätzer für FDR :

$$\widehat{FDR} = \hat{\pi}_0 \frac{|\{d_i^b | d_i^b \geq t_{k,\beta}\}| / B}{|\{d_i | d_i \geq t_{k,\beta}\}|}$$

(analog zur t-Statistik)

- Bestimmung der Signifikanzgrenzen und die Schätzung von FDR werden mit verschiedenen Δ wiederholt, bis sie zusammen ein möglichst gutes Testergebnis liefern.

SAM / Wilcoxon Test (SAM-Wilc)

- Vorteil : Verteilung der Nullhypothese ist bekannt.
- Algorithmus :
 1. Berechnung der Rangsumme W_i und Erstellung der Ordnungsstatistik $W_{(1)} \leq \dots \leq W_{(m)}$
 2. Berechnung der erwarteten Rangsummen $W_{(i)}^0$ aufgrund der $(i - 0.5)/m$ -Quantile.
 3. Bestimmung des ersten Datenpunktes ($W_{(i_1)}^0, W_{(i_1)}$) rechts des Mittelwertes $W_{mean} = n_1(n + 1)/2$, so dass $W_{(i)} - W_{(i)}^0 \geq \Delta$ bei vorgegebenem Δ ist.
 4. Analoge Schätzung von FDR :

$$\widehat{FDR}(\Delta) = \hat{\pi}_0 \frac{m(1 - \sum_{w=cut_{down}+1}^{cut_{up}-1} f_0(w))}{\#significantgenes}$$

(5. Wiederholung der Schritte 3. und 4.)

Empirical Bayes Analysis of Microarrays

- Die gemittelte Wahrscheinlichkeit für den Score von d_i über verschiedene Mittelwerte lautet für die Nullhypothese :

$$f_0(d_i) = \int \prod_{j=1}^I f_{obs}(x_{ij}|\mu)\pi(\mu)d\mu$$

- Bei differentiell exprimierten Genen gilt die Wahrscheinlichkeitsverteilung :

$$f_1(x_i) = f_0(d_{i1})f_0(d_{i2})$$

- Sind π_0 und π_1 die a priori Wahrscheinlichkeiten für die Nullhypothese, bzw. die Alternative, dann lässt sich die gesamte Wahrscheinlichkeitsverteilung wie folgt definieren :

$$f(d_i) = \pi_0 \cdot f_0(d_i) + \pi_1 \cdot f_1(d_i)$$

EBAM

- Die a posteriori Wahrscheinlichkeit für d_i lautet damit :

$$p_1(d_i) = 1 - \pi_0 \frac{f_0(d_i)}{f(d_i)}$$

- Der Schätzer für die FDR lautet wieder ganz analog :

$$\widehat{FDR} = \widehat{\pi}_0 \frac{|\{d_i^b | d_i^b \in \Gamma\}| / B}{|\{d_i | d_i \in \Gamma\}|}$$

$$\Gamma = \{d : p_1(d) \geq 0.9\}$$

- π_0 , f_0 und f müssen hier geschätzt werden, wobei die Bestimmung von f / f_0 mithilfe einer Regressionsanalyse ausreichend ist.

EBAM - Wilcoxon Test

- Wird die EBAM basierend auf dem Wilcoxon Test durchgeführt, ist die Verteilung f_0 unter der Nullhypothese bekannt. Folglich muss nur f geschätzt werden.
- Die a priori Wahrscheinlichkeit $\pi_0(\lambda)$ kann durch folgenden Ausdruck bestimmt werden :

$$\hat{\pi}_0(\lambda) = \frac{|\{p_i | p_i > \lambda\}|}{(1-\lambda)m}$$

Vergleich

- Die vier vorgestellten Statistikverfahren wurden auf drei Datensätze angewendet.
- In einer 5,000x50 Matrix wurden zufällige, normalverteilte Expressionswerte erzeugt. Auf die ersten 250 Gene wurde auf die ersten 25 Proben eine normalverteilte Störung mit positivem Mittelwert, auf die folgenden 250 Gene analog eine Störung mit negativem Mittelwert addiert.
- Die Genexpressionslevel von 3,226 Gene wurden insgesamt 15 mal für Zellen, die entweder Mutationen des BRCA1 oder das BRCA2 Gen tragen, mit Hilfe von cDNA Microarrays gemessen. Die beiden Mutationen sind hochgradig Krebsauslösend. (Hedenfalk-Daten)
- Mithilfe von Affymetrix high-density oligonucleotide chips wurden die Genexpressionsdaten von 3,051 Genen bei insgesamt 38 Patienten, die entweder an ALL oder an AML erkrankt waren, gemessen.

Vergleich

| Method | Simulation | | Hedenfalk | | Golub | |
|-----------|------------|------|-----------|------|----------|------|
| | <i>R</i> | FDR | <i>R</i> | FDR | <i>R</i> | FDR |
| SAM | 386.5 | 0.84 | 158 | 5.93 | 707 | 2.72 |
| SAM-Wilc | 369.1 | 0.88 | 206 | 7.25 | 714 | 2.75 |
| EBAM | 380.9 | 0.86 | 162 | 5.52 | 714 | 2.76 |
| EBAM-Wilc | 395.8 | 1.25 | 178 | 6.04 | 711 | 2.68 |

Zusammenfassung

- Die vier vorgestellten Teststatistiken beruhen entweder auf der t-Statistik oder der Wilcoxon Statistik
- Alle Tests machen sich die Eigenschaft zunutze, dass zwischen den Expressionswerten "implizite" Abhängigkeiten bestehen, nämlich dahingehend, dass die Anzahl der differentiell exprimierten Gene durch Betrachtung aller Gene abgeschätzt werden kann. Dieser Informationsgehalt erlaubt eine bessere Vorhersage über vorhandene differentielle Genexpression.
- Bei EBAM wird dieser Effekt dadurch ausgenutzt, dass die Wahrscheinlichkeit für einen Datensatz a posteriori berechnet wird.
- Bei SAM geht diese Information dadurch ein, dass die Nullhypothese durch viele Permutationen der Probenlabel geschätzt wird.
- Der Wilcoxon Statistik hat den Vorteil, dass die Verteilung der Nullhypothese exakt bekannt ist.

References

- [1] H. Schwender et al. : Comparison of the Empirical Bayes and the Significance Analysis of Microarrays, Promotionsarbeit
- [2] G. Parmigiani et al. : The Analysis of Gene Expression Data, Springer-Verlag, New York, 2003
- [3] U. Krengel : Einführung in die Wahrscheinlichkeits-theorie und Statistik, 6. Auflage, 2002, Vieweg