

Datamining

Bioinformatik 5. Semester

Werner Backfrieder

Backfrieder-Hagenberg

Data Mining

- Auswertung der DNA-Chips
 - Quantifizierte + normalisierte Spots
- Modellierung der Daten
 - Relation der Information aus Spots zu:
 - Biologischen Modellen
 - Genen
 - Genprodukten
 - Zellen
 - Gewebsanalysen
 - Kriminaltechnische Analysen

Backfrieder-Hagenberg

Methoden

- Skalierung
- Scatter Plot
- Principle Components Analyse (PCA)
- Cluster Analyse
- Self Organizing Maps (SOF)

Backfrieder-Hagenberg

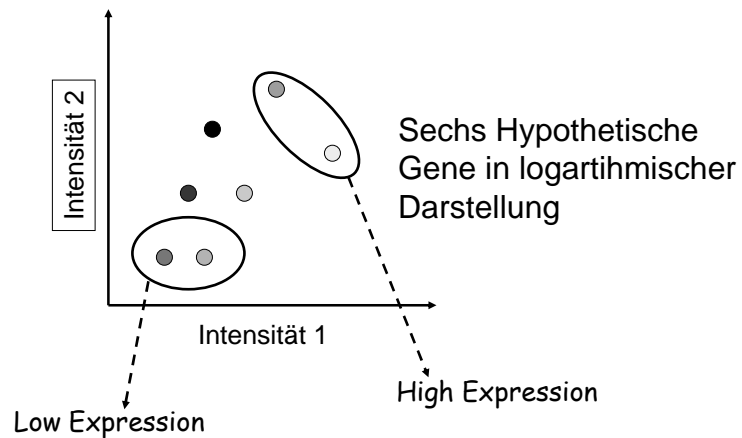
Skalierung

- Datenwerte 16 Bit [0,65535]
- Transformation mit 10er Logarithmus
 - $f(x) \rightarrow y: 10^y = x$
 - [1,65535] \longrightarrow [0,4.8]
- Vorteile
 - Einheitliche Verteilung
 - Analyse Dynamischer Prozesse (expression)
 - Ratios [-5,+5] (Faktor 100000!!!)

Backfrieder-Hagenberg

Scatter Plot

- Vergleich zweier Samples (Test vs. Referenz)



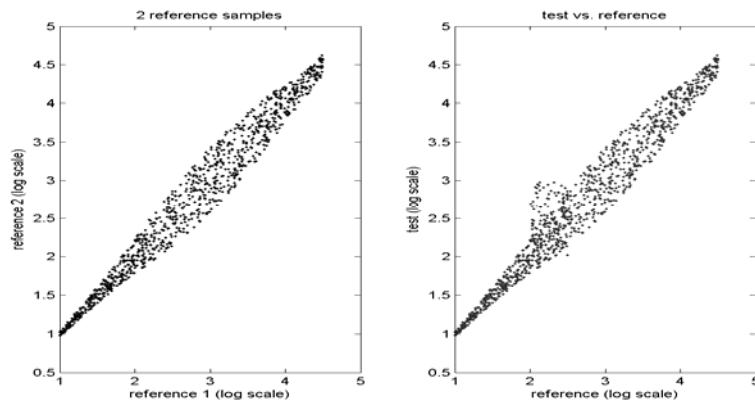
Backfriedler-Hagenberg

Scatter Plot

- Intuitive Darstellung zum Vergleich zweier Datensätze
- Identische Expressionen entlang der ersten Mediane
- Dynamische Expression (starke Veränderung) => starke Abweichung von erster Mediane
- Definierte Bedingungen: Vergleich zweier Chips mit gleicher Gen-Expressionen

Backfriedler-Hagenberg

Beispiele



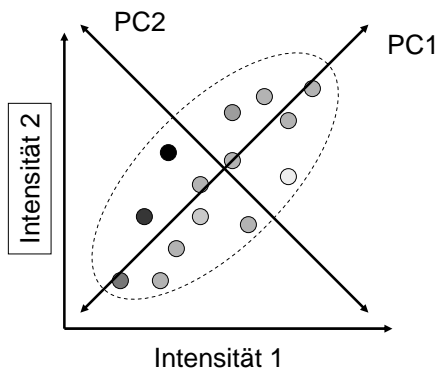
Backfrieder-Hagenberg

Principal Components Analyse (PCA)

- Vergleich mehrerer Chip-Datensätze
 - Scatter-Plot ungeeignet, z.B. in realem Experiment Vergleich mehrerer hundert Samples = Anzahl der Dimensionen
- Darstellung unmöglich
- => PCA Reduktion der Dimensionen
- PCA erhält Relationen aus höheren Dimensionen
- Multivariate Methode
- Häufigst verwendete Methode zur Analyse von DNA-Chips

Backfrieder-Hagenberg

PCA: Prinzipien



- Koordinaten-Trafo
- Ursprung im Mittelwert
- PC1 in Richtung größter Varianz
- Höhere Dimensionen normal auf alle anderen
- Unkorrelierte Variablen

Backfrieder-Hagenberg

PCA: Berechnung

- X_{ij} = Chip i , Spot j
 - I = No. Chips, J = No. Spots
- Kovarianz-Matrix C [$I \times I$]
- Diagonalisierung $C = U' V U$
 - U, V = Eigenvektoren, -werte
- Dimensionsreduktion
 - Auswahl der N höchsten Eigenwerte
 - Eigenvektoren Hauptachsen
 - Z-Scores: $z_i = U'_i * (X - X_m)$

Backfrieder-Hagenberg

Identitäten

1-dim

$$p(x) = (2\pi\sigma)^{-1/2} \exp\left(-0.5 \frac{(x-\mu)^2}{\sigma^2}\right)$$

$$\sigma^2 = \frac{\sum_i (x_i - \mu)^2}{n(n-1)}$$

$$z = \frac{(x - \mu)}{\sigma}$$

Multivariat

$$p(\vec{x}) = |2\pi\sigma|^{-N/2} \exp\left[(\vec{x} - \vec{\mu})' \sigma^{-1} (\vec{x} - \vec{\mu})\right]$$

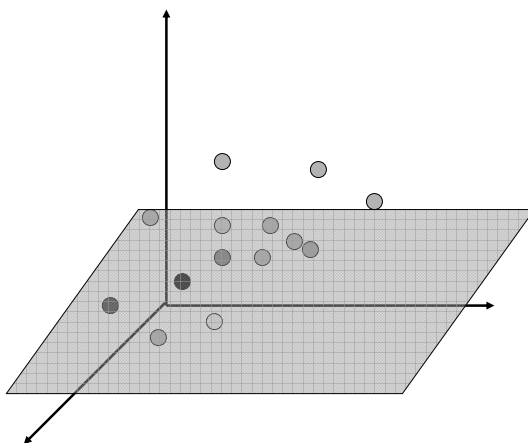
$$\sigma_{ij} = \frac{\sum_k (x_{ik} - \mu_k)(x_{jk} - \mu_k)}{n(n-1)}$$

$$\sigma = U' V U$$

$$z_i = U'_i (\vec{x} - \vec{\mu})$$

Backfrieder-Hagenberg

Graphische Interpretation



PCs spannen Ebene mit minimalem Abstand zu Punkten auf.

Regression

Dimensionsreduktion

Backfrieder-Hagenberg