Bernd Klaus (bernd.klaus@imise.uni-leipzig.de)
Verena Zuber (verena.zuber@imise.uni-leipzig.de)

http://uni-leipzig.de/~zuber/teaching/ws10/r-kurs/

Auch dieses Übungsblatt ist dem Datensatz aus der Veröffentlichung *Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression* von Golub et al. gewidmet.

- Genexpressionsdaten mit $n = 38$ Beobachtungen und $p = 3051$ Genen
- Faktorvariable `golub.cl` beschreibt welche Form von Leukämie ($k = 2$) bei der entsprechenden Beobachtungen vorliegt
  - ALL: acute lymphoblastic leukemia ($n_0 = 27$)
  - AML: acute myeloid leukemia ($n_1 = 11$)
- Die Matrix `golub.gnames` enthält Information zu den Bezeichnungen der beobachteten Gene
- Die Daten sind schon praeprozessiert und zu finden als `RData`-File im Netzwerkordner unter L:\R-Kurs oder im `R`-Paket `multtest` unter *Data*

1 Aufgabe: Differentielle Expression

(a) Berechnen Sie für jedes Gen den Fold Change (`ALL-AML`).

(b) Wie viele Gene sind differentiell unterexprimiert? Erstellen Sie eine Liste mit den differentiell unterexprimierten Genen (negativer Fold Change).

(c) Berechnen Sie für jedes Gen den t-Score. Benutzen Sie dazu die Varianzen aus Aufgabe 1 der letzten Woche.

(d) Erstellen Sie zwei Ranglisten einmal nach dem Fold-Change einmal nach dem t-Score. Vergleichen Sie die Top20 der beiden Listen. Welche Gene tauchen in beiden auf?

(HINWEIS: Nutzen Sie zur Lösung die `match` Funktion.)

(e) Erstellen Sie ein Histogramm der t-Scores. Beschreiben Sie die Form.

(f) Berechnen Sie den Mittelwert und die Standardabweichung für die 90% der mittleren absoluten t-Scores. Fügen Sie in das Histogramm eine Normalverteilung mit diesen Parametern ein. Was fällt Ihnen insbesondere bei den niedrigsten und höchsten t-Scores auf?

2 Aufgabe: GeneNet

(a) Installieren Sie alle Pakete, die sich unter H:\befinden. Fügen Sie dazu zunächst dieses Verzeichnis zu den `.libPaths()` hinzu.

(b) Speichern Sie die Messwerte 250 Gene mit dem höchsten absoluten t-score in einer neuen Matrix ab. Fügen sie die Namen der Gene der Matrix hinzu und transponieren Sie diese. (So dass die Matrix 250 Spalten besitzt, jede Spalte also die jeweils 38 Expressionswerte eines Gens enthält.)

(c) Berechnen Sie mittels `GeneNet` die partiellen Korrelationen zwischen diesen Genen.

(d) Erstellen Sie ein Netzwerk, dass die 100 besten Knoten enthält (bezogen auf den fdr-Wert).