Wintersemester 2007/2008 Abgabe am 09. Januar 2008
Prof. Dr. Stefan Posch, Dipl.-Bioinf. Jan Grau

10. Übung „Angewandte Bioinformatik mit Perl und R“

1. Schauen Sie sich die Datei `golub-data.rdb` an. Sie enthält je Zeile die mRNA-Expressionwerte eines Genes, und zwar für Gewebeproben zweier unterschiedlicher Leukämie-Typen, der akute lymphoblastische Leukämie (all) und akute myeloische Leukämie (aml). Die erste Spalte gibt eine Accession für das jeweilige Gene an.

   Lesen Sie diese Tabelle in einen `data frame` ein. Stellen Sie in R fest, wieviele Gene und wieviele Proben beider Leukämietypen in dem Datensatz vorhanden sind.

   Probieren Sie das Kommando:

   ```
   golub <- read.table( "golub-data.rdb",
                sep = "\t", quote = "", header=TRUE, row.names = 1)
   ```

   und schauen Sie sich das Resultat an. Lesen Sie in der Dokumentation von `read.table` die Bedeutung nach.

   (1 Punkt)

2. In der Datei `golub-description.rdb` sind zu den Gene-Accessions Beschreibungen enthalten. Lesen Sie auch diese Datei ein. Hinweis: Geben Sie als zusätzliches Argument für `read.table` an: `as.is=TRUE`, also

   ```
   .... read.table( ......., as.is=TRUE)
   ```

   Überzeugen Sie sich, dass die Accessions in beiden Dateien identisch und in der gleichen Reihenfolge sind.

   (1 Punkt)

3. Benutzen die Funktion `grep`, um die Expressionswerte aller Gene zu extrahieren, die gemäß Beschreibung Kinasen sind. Wieviele gibt es?

   Plotten Sie die Expressionswerts aller Kinasen für die Probe `all3`. Schreiben Sie eine Funktion, die den Namen einer Probe und einen Dataframe nimmt, und die zugehörigen Werte plottet. Als Label an der y-Achse soll der Name der Probe erscheinen.

   Nutzen Sie diese Funktion und plotten Sie die Expressionswerte aller Kinasen der ersten vier `aml`-Proben in jeweils einem eigenen Plot aber einer einzigen Graphik. Nun für dieselben vier Proben in einem gemeinsamen Plot. Verschönern Sie mit Farben, Legenden, etc.

   Lassen Sie (in einer Graphik) die boxplots der Proben `all1` und `all3` zeichnen.

   (4 Punkte)

4. Selektieren Sie nun alle Expressionswerte eines der Gene (über alle Leukämieproben), also eine Zeile aus dem `data frame`. Wenn Sie davon z.B. den Mittelwert oder Varianz bestimmen wollen, werden Sie eine Fehlermeldung bekommen. Dies liegt daran, dass ein `data frame` eine Liste ist, und auch eine selektierte Zeile eine Liste. Mit der Funktion `as.matrix` kann man daraus einen Vektor (oder bei mehreren selektierten Zeilen eine Matrix) machen.

   Plotten Sie die Expressionswerte so, dass die zwei Gruppen gut unterscheidbar sind.

   Erzeugen Sie sich zwei Vektoren, mit denen Sie die all- bzw. aml-Proben bequem selektieren können. Machen Sie das so, dass es auch bei anderer Anzahl oder Reihenfolge der Proben funktioniert.

   Selektieren Sie für ein beliebiges Gen nun die Expressionswerte der Leukämietypen in zwei verschiedene Vektoren. Testen Sie per t-Test, ob die mittlere Expression in den beiden Gruppen (signifikant) unterschiedliche ist.

   Schreiben Sie hierfür eine Funktion, die diesen t-Test für ein beliebiges Gen durchführt. Wenden Sie Funktion auf alle Gene an (oder soviel Gene, wie Ihr Rechner zuläßt) und speichern Sie p-Values in einem Vektor. Sortieren Sie die p-Values mit der Funktion `sort` und bestimmen Sie damit das Gen mit dem kleinsten p-Value.

   Können Sie etwas über das Gen herausfinden?

   (4 Punkte)