Wintersemester 2007/2008 Abgabe am 16. Januar 2008
Prof. Dr. Stefan Posch, Dipl.-Bioinf. Jan Grau

11. Übung „Angewandte Bioinformatik mit Perl und R“

1. Wir betrachten zunächst noch einmal die `golub` Daten aus der letzten Übung:
   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?
   (3 Punkte)

2. Im Linux-Pool ist Bioconductor unter `/lehre/agprbio/Biocond/lib` installiert. Um Pakete mittels der Funktion `library()` von diesen Nicht-Standard-Pfad zu laden, können Sie z.B. in Ihrer shell die Environment-Variable `R_LIBS` entsprechend setzen, in der `tcsh` also

   ```
   setenv R_LIBS /lehre/agprbio/Biocond/lib
   ```

   setzen.

   Sie können auch in R den Aufruf
   `.libPaths("/lehre/agprbio/Biocond/lib")`
   verwenden.

   Für die folgenden Aufgaben können Sie die Folien des Bioconductor nutzen, diese sind unter `www.bioconductor.org` und auch auf der Web-Seite der Vorlesung zu finden.
   (0 Punkte)

3. Lesen Sie den Dilution Datensatz ein. Wieviele Gene enthält jedes array? Wieviele und welche Gen-Namen enthalten als Teilstring “Lys”. Wieviele probes enthalten die zugehörigen probesets, geben Sie textuell deren Intensitäten aus.

   (2 Punkte)

4. Erstellen Sie ein Histogramm über die Größe der probesets, d.h. wieviele probesets gibt es für die verschiedenen Größen von probesets. Da eine probeset-Größe sehr oft vorkommt, können wir in diesem Histogramm nicht viel sehen. Überlegen (und realisieren) Sie eine bessere Darstellung Welche Größe kommt am häufigsten vor, wie oft?

   Welches/welche Gen/Gene hat/haben die größte Anzahl an probes in ihrem probeset? Plotten Sie die Intenistäten und versuchen Sie möglichst viel Informationen über das/die Gen/Gene zu finden.

   (3 Punkte)

5. Plotten Sie für alle Gene mit dem Teilstring “Lys” im Namen die Intensitäten über alle probes, und zwar einmal die perfect matches, dann die mis-matches

   (2 Punkte)