

Übungen zur Vorlesung
**Einführung in die angewandte
Bioinformatik**
Sommersemester 2009

Übungsblatt 11
Bearbeitungszeit:
09.07.2009

Notieren Sie bitte auf diesem Übungsblatt auch immer die R-Befehle, die Sie zum Lösen einer Aufgabe benutzt haben.

Aufgabe 11.1 – Microarray-Daten im GEO finden

Gehen Sie auf die Seite des Gene Expression Omnibus und suchen Sie nach einem Datensatz, mit dem der Effekt der Gabe von Interleukin 13 auf Bronchialzellen mit einer Zeitreihe („time course“) von Experimenten untersucht wird. Wie lautet die Zugriffsnummer des Datensatzes?

Um welchen Microarray-Typ handelt es sich?

Betrachten Sie die Samples GSM47468 und GSM47459 genauer. Worin unterscheiden Sie sich?

Sie können für jedes Sample auch unter „View full table“ an die komplette Liste aller Intensitäten herankommen, aber leider kann die so erhaltene Datei nicht direkt von R verarbeitet werden und muss erst angepasst werden. Diesen Schritt haben wir Ihnen abgenommen. Laden Sie daher die Dateien gsm47459.txt und gsm47468.txt von der Übungswebseite herunter.

Starten Sie nun R und lesen Sie die beiden Dateien in zwei Variablen ein. Die Benennung dieser Variablen bleibt Ihnen überlassen, aber im Folgenden verwenden wir x für GSM47459 und y für GSM47468.

Wie sind die Dimensionen und wie lauten die Spaltennamen der Datensätze?

Aufgabe 11.2 – Boxplot

Erstellen Sie einen Boxplot der value-Spalte des x-Datensatzes.

Tip: Um in R z. B. die value-Spalte von x anzusprechen, müssen Sie normalerweise x\$value schreiben. Sie können dies aber auch mit x\$v abkürzen, da „v“ eindeutig ist – nur Spalte „value“ des Datensatzes kann gemeint sein.

Erstellen Sie nun zwei weitere separate Boxplots dieses Datensatzes, einer soll nur die Werte enthalten, die in der abs-Spalte als „P“ markiert sind und einer nur die, die als „A“ markiert sind. Stellen Sie die y-Achse logarithmiert dar, indem Sie den Parameter log="y" setzen.

```
> boxplot(x$value[x$abs == "P"], log="y")  
> boxplot(x$value[x$abs == "A"], log="y")
```

Wer möchte, kann zwischen beiden Befehlen noch `x11()` ausführen, um ein neues Plotfenster zu bekommen.

Versuchen Sie, beide Boxplots mit nur einer Befehlszeile innerhalb eines einzigen Fensters anzuzeigen (wieder logarithmiert).

Wo liegen (grob geschätzt) jeweils die beiden mittleren Querbalken?

Wie könnten Sie den Wert exakt ausrechnen?

Aufgabe 11.3 – Verteilung der P-Intensitäten

Suchen Sie sich einen der beiden Datensätze (x oder y) aus. Sind die Logarithmen der P-Intensitäten normalverteilt? (Das geht z. B. mit einem Q-Q-Plot.)

Sind auch die nicht-logarithmierten Werte normalverteilt?

Aufgabe 11.4 – Scatterplot

Plotten Sie nun die Werte der beiden Datensätze gegeneinander in einem Scatterplot. Beide Achsen sollen logarithmiert dargestellt werden, dies geht mit dem Parameter `log="xy"`.

Erstellen Sie folgendermaßen einen Vektor, der markiert, welche Datenpunkte in x als P, aber in y als A markiert sind:

```
> p_a = (x$abs=="P" & y$abs=="A")      # & bedeutet "und"
> p_a
```

Zeichnen Sie diese Punkte von `x$val` nun in blau in Ihren bestehenden Plot ein. Benutzen Sie dazu den `points`-Befehl. Er funktioniert ähnlich wie `plot`, löscht aber einen eventuell bereits angezeigten Plot nicht. Zeichnen Sie auch die anderen Kombinationen (A/P, P/P, A/A) in jeweils anderen Farben ein.

Wo liegen hauptsächlich die als A/A markierten Punkte?

Aufgabe 11.5 – Eine differenzielle Expression finden

Sie möchten nun schauen, welcher der Punkte am stärksten von der Diagonalen abweicht. Dazu nehmen Sie das Verhältnis von x- zu y-Intensität, aber um zu vermeiden, dass geringe Intensitäten bevorzugt werden, addieren Sie vorher 2000: $\frac{x$val+2000}{y$val+2000}$

Finden Sie mit `which.max` den Index des Punktes, bei dem dieses Verhältnis am größten ist.

Zeichnen Sie diesen Punkt in einer weiteren Farbe in Ihren Plot. Verwenden Sie den Parameter `pch="x"`, um die Form des Punktes zu ändern. Ermitteln Sie nun noch, welchen Namen (id) der Spot im Datensatz hat.

Aufgabe 11.6 – MA-Plot

Plotten Sie die Intensitäten der beiden Datensätze in einem MA-Plot. Zur Erinnerung:

$$A = 0.5(\log x_{val} + \log y_{val})$$

$$M = \log x_{val} - \log y_{val}$$

Nehmen Sie A als x-Achse und M als y-Achse.

Aufgabe 11.7 – Ein passendes Gen finden

Nun geht es darum, den in der vorletzten Aufgabe gefundenen Namen wieder einem Gen zuzuordnen. Gehen Sie dazu wieder zurück auf die Webseite zu dem Interleukin-13-Datensatz. Dort findet sich hinter „Platform“ ein Link („GPL96“), hinter dem sich u. A. eine Beschreibung der auf dem Array verwendeten Oligonukleotide verbirgt. Laden Sie die volle Tabelle herunter, öffnen Sie sie in einem Texteditor und suchen Sie nach der ID, die Sie aus Ihren Daten wissen.

Sie finden in der Zeile der Datei mehrere Informationen. Suchen Sie damit nach dem passenden Gen in der Uniprot-Datenbank. Wie heißt das Gen? _____

Wie lautet die Uniprot-Zugriffsnummer des ersten Treffers?

Auf welchem Chromosom liegt das Gen (folgen Sie dazu einem geeigneten Querverweis)?

Wie schwer ist das entsprechende Protein? _____