

Übungen zur Vorlesung Algorithmische Bioinformatik

Freie Universität Berlin, WS 2008/09

Roland Krause · Clemens Gröpl · Hannes Luz · Utz J. Pape · Martin Vingron

Blatt 10 · Ausgabe am 20.01.2009

Abgabe am 27.01.2009 vor Beginn der Vorlesung

Aufgabe 41 (Lowess Normalisierung). Ziel der Aufgabe ist es, sich die MA plots von einem Experiment anzuschauen, anschliessend eine lowess Normalisierung durchzuführen und das Ergebnis zu betrachten.

1. Lade die Daten `wood_data.txt` von der Website und lies sie in R ein. Das geht ganz gut mit `read.delim("wood_data.txt")`. Beachte, dass die ersten beiden Spalten Gen- und Reportername enthalten. Die weiteren Spalten sind nach der Array ID benannt und dem entsprechenden Farbkanal (R und G).
2. Erstelle für jeden Array einen MA-Plot inkl. einer Geraden für den Median. Dabei kann es passieren, dass für einzelne Gene ungültige Werte herausbekommt. Warum? Am besten man ignoriert diese einfach. Wenn man sich die MA Plots anschaut, was bemerkt man?
3. Führe eine lowess Normalisierung (R Befehl `lowess`) für alle Arrays durch. Experimentiere dabei mit der Fenstergrösse. Wie macht sich der Einfluss bemerkbar? Stelle dafür sowohl die nichtnormalisierten Daten mit der Regressionslinie als auch die normalisierten Daten dar.
4. *Freiwilliger Zusatz: Verwende verschiedene Kernel-Functions in der lowess Normalisierung und interpretiere das Ergebnis.*

Aufgabe 42 (Quantilnormalisierung). Eine Möglichkeit zur Normalisierung von Daten ist die Quantil-Normalisierung.

1. Ziehe 1000 Sample aus einer Normalverteilung mit Erwartungswert 1 und Varianz 1, sowie 1000 Sample aus einer γ -Verteilung mit shape gleich 2 und scale gleich 1.
2. Erzeuge qq-plots gegen eine Standard-Normalverteilung und der beiden Verteilungen gegeneinander. Was beobachtest Du? Warum?
3. Benutze die Quantil-Normalisierung, um beide Verteilungen miteinander zu normalisieren. Erstelle noch einmal einen qq-plot gegen eine Standard-Normalverteilung. Was beobachtest Du? Was kann man daraus für eine Gefahr dieser Normalisierung schliessen? Warum macht es keinen Sinn, die beiden normalisierten Daten gegeneinander zu plotten (mit einem qq-plot)? Was würdest Du beobachten?

Aufgabe 43 (Explorative Analyse und Normalisierung von Microarrays). Ein erster Schritt bei der Auswertung von Microarrays ist es stets, sich die Daten zunächst anzuschauen (explorative Datenanalyse), um sich anschliessend Gedanken über die Normalisierung zu machen und diese dann durchzuführen. Das ist der Inhalt dieser Aufgabe. Die Aufgabe sollte in R mit installiertem `Biobase` ausgeführt werden, da dort schon die meisten Funktionen implementiert sind. Ansonsten würde es sehr viel Arbeit werden.

1. Benutze den `lymphoma` Datensatz, um eine kurze explorative Datenanalyse durchzuführen. Der Datensatz ist Teil des `vsnp` Pakets und wird folgendermaßen eingelesen:

```
library(vsn);data(lymphoma);e <- exprs(lymphoma)
```

Dabei enthält `e` am Ende die Expressionsmatrix. Es genügt, wenn Ihr Euch 4 Chips aussucht und diese analysiert. Für die explorative Analyse sind folgende Werkzeuge hilfreich, die auf die logarithmierten Daten (Gene mit ungültigen Daten wieder raus-schmeißen!) angewandt werden sollten:

- Verteilung der Expressionswerte pro Chip (boxplot, qq-plot)
- Scatterplots: Führe dies sowohl für die logarithmisch transformierten Daten wie auch für die noch nicht transformierten Daten durch und vergleiche.
- Matrix of Absolute Deviation (MAD)
- Plot der Standardabweichung gegen den Mittelwert (sdm plots), dafür eignet sich der Befehl `meanSdPlot`. Führe dies sowohl für die logarithmisch transformierten Daten wie auch für die noch nicht transformierten Daten durch und vergleiche.
- *Freiwillig für ganz arbeitswütige: Korrespondenzanalyse*

2. Normalisierung und Vergleich

- (a) Was ist die wichtigste Annahme, um eine Normalisierung z.B. mit `vsnp` durchzuführen? Warum ist das wichtig?
- (b) Verwende `vsnp` zur Normalisierung der Daten (keine Hintergrundkorrektur und Summarizing mit `median polish`). Das muss man nicht selbst implementieren, sondern man kann einfach das Paket `vsnp` in R benutzen.
- (c) Vergleiche die transformierten Daten mit den originalen Daten. Wie soll der Vergleich aussehen? Boxplots, Scatterplots zwischen den normalisierten Daten und diese Plots vergleichen mit den Rohdaten und den logarithmierten Daten. Gleiches auch für `sdm` Plots.

3. *Und wer noch immer nicht genug hat: Benutze eine andere Normalisierungsmethode Deiner Wahl (z.B. RMA für Hintergrundkorrektur, Quantilnormalisierung und zum Zusammenfassen `median polish`; oder probiere es mit `MAS5`). Vergleiche das Ergebnis mit den `vsnp` transformierten Daten und mit den originalen Daten. Was stellst Du fest?*