

Übungen zur Vorlesung Algorithmische Bioinformatik

Freie Universität Berlin, WS 2016/17

Martin Vingron · Annalisa Marsico · Alena van Bömmel · Edgar Steiger · Thimo Wellner

Blatt 11 · Ausgabe am 9.1.2017

Abgabe am 16.1.2017 vor Beginn der Vorlesung

Name:

Matrikelnummer:

Übungsgruppe:

Aufgabe 1 (10 Punkte; Theorie). Gegeben seien folgende Expressionsdaten von Genen ($G1, \dots, G7$) in drei verschiedenen Experimenten (A, B und C):

	A	B	C
G1	20	1	2
G2	19.5	1	1
G3	19	0.5	1
G4	22.5	3	6
G5	18	1	0
G6	19	0	0
G7	20.5	1.5	3

Normalisieren Sie die Experimente mit Hilfe der Quantil-Normalisierung, wie sie in der Vorlesung vorgestellt wurde.

Aufgabe 2 (30 Punkte; Praxis). Wir möchten die Expressionsdaten¹ von Krebspatienten (K) und gesunden Patienten (G) analysieren. Machen Sie sich zuerst mit den Daten vertraut:

1. Laden Sie die Tabelle in R und schauen Sie sich die Statistiken der Expressionswerte in den verschiedenen Experimenten an (z.B. *summary*, *boxplot*, *hist*, *str*, *mean*, ...). Was fällt Ihnen auf? Welche Transformation der Daten würden Sie vorschlagen?
2. Vergleichen Sie die Expressionswerte von Experiment 3 mit Experiment 8 (G_{20} vs. G_{25}), sowie Experiment 1 mit Experiment 8 (K_{18} vs. G_{25}). Erstellen Sie dazu jeweils einen Plot (z.B. mit *smoothScatter*), in dem Sie die logarithmierten Expressionswerte des jeweiligen Experiments gegeneinander auftragen. Was beobachten Sie?
3. Erstellen Sie einen MA-Plot von den Expressionswerten. Führen Sie dazu folgende Berechnung durch:

$$M_i = \log_2(K_i/G_i); A = 0.5 \cdot \log_2(K_i G_i),$$

wobei für jedes Gen i der Expressionswert K_i im Krebspatienten K_{18} und G_i der Expressionswert im gesunden Patienten G_{25} ist. Was beobachten Sie?

¹Material 1: https://www.molgen.mpg.de/3722074/expr_cel.txt

Aufgabe 3 (20 Punkte; Programmieren). Wir möchten nun die Experimente aus Aufgabe 2 normalisieren.

Schreiben Sie dafür ein Programm, welches

1. eine Expressionsmatrix und die zugehörige Unterteilung der Patienten in gesund vs. erkrankt aus einer Datei einliest. Der Dateiname wird als Parameter in der Kommandozeile übergeben.
2. eine Funktion enthält, die die eingelesenen Daten als Input erhält, die Expressionswerte logarithmiert (\log_{10}) und auf diesen die Quantil-Normalisierung ausführt. Implementieren Sie dazu die Quantil-Normalisierung, wie sie in der Vorlesung vorgestellt wurde. Verwenden Sie **keine** implementierten Funktionen wie z.B. *normalize.quantiles* oder ähnliches.
3. die normalisierte Matrix in eine Datei, deren Name als 2. Parameter in der Kommandozeile übergeben wird, ausschreibt.
4. Führen Sie Ihr Programm auf den gegebenen Daten aus und berechnen Sie erneut die Statistiken der Expressionswerte wie in Aufgabe 2 für jedes Experiment. Was fällt Ihnen auf?

Aufgabe 4 (40 Punkte; Programmieren). Wir möchten nun differentiell exprimierte Gene zwischen den zwei Patientengruppen finden.

1. Ergänzen Sie Ihr Programm aus Aufgabe 3 und implementieren Sie eine Funktion, die als Input die gegebenen Daten erhält, die t -Statistik und den zugehörigen p -Wert berechnet. Sie können gleiche Varianz in der beiden Gruppen annehmen und die *degrees of freedom* wie folgt berechnen: $df = n_1 + n_2 - 2$, wo n_i die Anzahl der Experimente in der jeweiligen Gruppe bezeichnet.
2. Führen Sie Ihre Funktion auf den normalisierten Daten aus. Welches sind die top-10 hochregulierten Gene in den Krebs- bzw. in den gesunden Patienten?
3. Wie viele signifikant differentiell exprimierte Gene finden Sie auf dem Signifikanzniveau von $\alpha = 0.05$? Begründen Sie kurz, warum die Anzahl der so gefundenen signifikanten Gene zu hoch ist.
4. Führen Sie nun die Benjamini-Hochberg-Korrektur auf den p -Werten aus dem t -Test durch. Implementieren Sie dazu eine Funktion, die die Korrektur mit gegebenem Parameter α berechnet und die Signifikanz der Einträge als SIGNIFIKANT/NICHT SIGNIFIKANT ausgibt. Verwenden Sie den in der Vorlesung beschriebenen Algorithmus und verwenden Sie **keine** implementierten Funktionen wie *p.adjust* oder ähnliches.
5. Wie viele signifikanten Gene finden Sie nach der Korrektur?