Übungen zur Vorlesung Algorithmische Bioinformatik

Freie Universität Berlin, WS 2014/15
Martin Vingron · Juliane Perner · Annkatrin Bressin

Blatt 10 · Ausgabe am 15.12.2014
Abgabe am 05.01.2014 vor Beginn der Vorlesung

Name: Matrikelnummer: Übungsgruppe:

Aufgabe 1 (*40 Punkte; Praxis*)**.** Wir möchten die Expressionsdaten[1] 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. mit *summary* auf jeder Spalte und/oder *boxplot*). Was fällt Ihnen auf?

2. Führen Sie nun ein hierarchisches Clustering auf den Experimenten aus und stellen Sie das Dendrogram grafisch dar. Dazu benötigen Sie die Funktionen *dist*, *hclust* und *plot*. Logarithmieren Sie vorher die Daten und wählen Sie ein geeignetes Distanzmaß und eine geeignete Clustering-Methode aus. Finden Sie alle Replikate dort, wo Sie sie erwarten würden? Begründen Sie ihre Antwort.

3. Vergleichen Sie die Expressionswerte von Experiment 3 mit Experiment 8 (G_{20} vs. G_{25}), so wie Experiment 1 mit Experiment 8 (K_{18} vs. G_{25}). Erstellen Sie dazu einen Plot (z.B mit *smoothScatter*) in dem Sie die logarithmierten Expressionswerte des jeweiligen Experiments gegeneinander auftragen. Was beobachten Sie?

4. Erstellen Sie einen MA-Plot von den Expressionswerten. Führen Sie dazu folgende Berechnung durch:
$$M_i = log_2(K_i/G_i); A = 0.5 log_2(K_i G_i)$$
Wobei für jedes Gen i, K_i der Expressionswert im Krebspatienten K_{18} und G_i der Expressionswert im gesunden Patienten G_{25} ist. Was beobachten Sie?

Aufgabe 2 (*25 Punkte; Programmieren*)**.** Wir möchten nun die Experimente aus Aufgabe 1 normalisieren.

1. Schreiben Sie ein Programm, das eine Expressionsmatrix und die zugehörige Unterteilung der Patienten in gesund vs. erkrankt aus einer Datei einliest. Schreiben Sie ihr Programm so, dass der Dateiname als Parameter in der Kommandozeile übergeben wird.

2. Implementieren Sie eine Funktion, die die eingelesenen Daten als Input erhält und auf diesen die Quantile-Normalisierung ausführt. Implementieren Sie dazu die Quantile-Normalisierung wie sie in der Vorlesung vorgestellt wurde.

3. Abschließend soll das Programm die normalisierte Matrix in eine Datei, deren Name als 2. Parameter in der Kommandozeile übergeben wird, ausschreiben.

[1] Material 1: `https://ws.molgen.mpg.de/ws/111250/expr_CEL.txt` **(verfügbar bis 29.12.2014)**

4. Führen Sie ihr Programm auf den gegebenen Daten aus und berechnen Sie erneut die Statistik der Expressionswerte für jedes Experiment. Was fällt Ihnen auf?

Aufgabe 3 (*35 Punkte; Programmieren*)**.** Wir möchten nun differentiell exprimierte Gene zwischen den zwei Patientengruppen finden.

1. Ergänzen Sie Ihr Programm aus Aufgabe 2 und implementieren Sie eine Funktion, die als Input die gegeben Daten erhält und die T-Statistik berechnet.

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. Bestimmen Sie die Funktion der differentiell exprimierten Gene. Nutzen Sie das online GeneID-conversion Tool[2].

4. Vergleichen Sie in einem Scatterplot die durchschnittlichen Expressionswerte der zwei Patientengruppen und markieren Sie die differentiell exprimierten Gene. Was fällt Ihnen auf?

[2] http://david.abcc.ncifcrf.gov/conversion.jsp

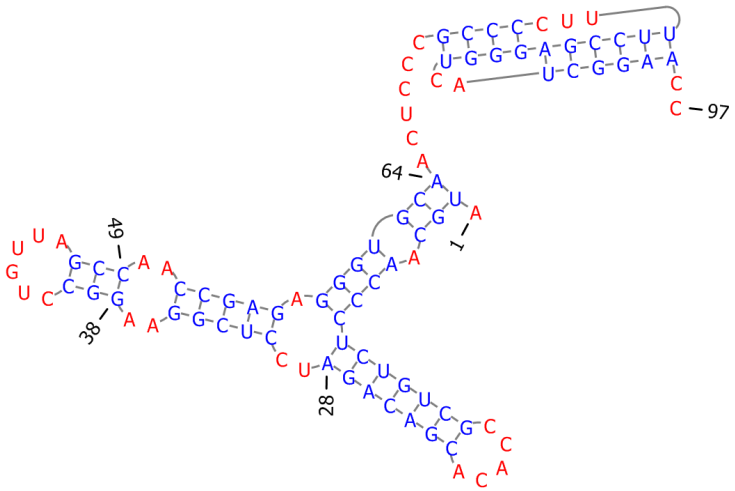
Aufgabe 4 (Optional)**.** Sie haben folgendes Sequenzalignment für die Bindestellen eines Transkriptionsfaktors erhalten:

CACA
TATA
TATC
TACA
CACT
TACC
CACA
TATG
AATT
TACA

1. Stellen Sie die Count matrix auf und transformieren Sie diese dann in eine positionsspezifische Log-Odds Score Matrix. Nehmen Sie dazu eine Hintergrundverteilung von $(0.3, 0.2, 0.2, 0.3)$ auf (A, C, G, T) an.

2. Berechnen Sie die relative Entropy jeder Position im Motif zur Hintergrundverteilung.

Aufgabe 5 (Optional)**.** Gegeben sei die folgende RNA-Sekundärstruktur.



1. Benennen Sie die darin enthaltenen Sekundärstrukturelemente. Kann diese Struktur durch den Nussinov-Algorithmus vorhergesagt werden? Begründen Sie ihre Antwort.

2. Formulieren Sie den Algorithmus zur RNA-Faltung nach Nussinov. Beschreiben Sie dazu den Aufbau der Score-Matrix in Pseudocode. Welche Laufzeit hat der Algorithmus?

Aufgabe 6 (Optional)**.** In folgender Tabelle sehen Sie, welche Sonden (*probes*, in Spalten) zu welchen Klonen (Zeilen) bei einem STS Content Mapping hybridisierten. Nun soll die Reihenfolge der Sonden bestimmt werden.

$Klon \backslash Sonde$	A	B	C	D
1	1	1	1	0
2	1	0	1	0
3	0	1	0	1
4	0	0	1	1

1. Berechnen Sie die optimale Permutation der Sonden indem Sie das Problem auf ein TSP reduzieren.
2. Welche Eigenschaft sollte die Matrix nun erfüllen? Welcher Eintrag in der ursprünglichen Tabelle rührt wahrscheinlich von einem Fehler her, wenn Ihre Lösung korrekt ist?

Aufgabe 7 (Optional)**.** Die Anzahl der Kunden, die ein Buchgeschäft betreten, lässt sich mit einem Poissonprozess mit der Intensität $\lambda = 4$ pro Stunde modellieren. Das Geschäft öffnet um 9:00 früh.

1. Wie groß ist die Wahrscheinlichkeit, dass genau 1 Kunde bis 9:30 ankommt?
2. Wie groß ist dann die Wahrscheinlichkeit, dass 5 Kunden bis 11:30 ankommen?
3. Wie groß ist die Wahrscheinlichkeit, dass insgesamt 10 Kunden an dem Tag das Geschäft betreten haben? Das Geschäft schliesst um 17 Uhr.