

Übungen zur Vorlesung Algorithmische Bioinformatik

Freie Universität Berlin, WS 2008/09

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

Blatt 11 · Ausgabe am 28.1.2008

Abgabe am 4.2.2008 vor Beginn der Vorlesung

Aufgabe 44 (Differentielle Gen-Analyse von Microarrays). Als Bioinformatiker kommt man hoffentlich auch mal zu dem Vergnügen, selbst neue Statistiken aufzustellen. Angenommen, Du hast soeben den t-Test erfunden und möchtest zeigen, dass er gut funktioniert, um differentiell exprimierte Gene zu detektieren. Genau das ist nun Deine Aufgabe unter der Berücksichtigung, dass Du nur begrenzte Ressourcen zur Verfügung hast. Zunächst simulierst Du Genexpressions-Daten für zwei Samples (z.B. krank und gesund), bei denen Du für jedes Gen weißt, ob es differentiell exprimiert ist oder nicht. Anschließend berechnest Du die t-Statistik für jedes der Gene und zeichnest die ROC-Kurve. Außerdem interessiert Dich der Einfluss der Anzahl der Replikate. Daher führst Du dies alles für je für 2,3 und 4 Replikate durch und vergleichst anschliessend die Ergebnisse.

Jetzt noch schrittweise. Die nachfolgende Anweisung ist für 2,3 und 4 Replikate getrennt durchzuführen.

1. Betrachte 1000 Gene von denen 100 Gene differentiell exprimiert sind.
 - Für die nicht-differentiellen Gene, ziehe die Expressionswerte (auch für die Replikate) aus einer Normalverteilung mit Erwartungswert 10 und Standardabweichung 5. Gehe so für beide Samples vor!
 - Die differentiellen Genexpressionswerte sind etwas komplizierter. Generiere zunächst auf gleiche Weise die Expressionswerte für das gesunde Sample. Die Expressionswerte für das kranke Sample erhältst Du, indem Du für jedes Replikat und Gene jeweils einen zufälligen Wert aus einer Normalverteilung mit Erwartungswert 5 und Standardabweichung 1 addierst.
2. Wende auf jedes Gen den t-Test an. Überlege Dir, ob Du einen gepaarten Test verwenden solltest. Plote die p -values auf sinnvolle Art.
3. Berechne für jeden möglichen Threshold wieviele differentielle Gene Du als differentiell klassifizierst relativ zu allen differentiellen Genen. Ausserdem brauchen wir noch die (komplementäre) Spezifität.

Und nun zur gemeinsamen Analyse der Replikate:

1. Plote die drei ROC-Kurven in einen Plot. Beschreibe und interpretiere was Du siehst!
2. Wie würde sich eine Veränderung der Erwartungswerte und Standardabweichungen bei der Simulation auswirken?
3. Welche grundsätzlichen Probleme siehst Du bei der Simulation?

4. Korrigiere gegen multiples Testen! Beschreibe Dein Vorgehen und die Ergebnisse und interpretiere!
5. Was müsstest Du zusätzlich machen, wenn Du Deine Ergebnisse publizieren wolltest, aber kein Labor zur Verfügung hast?
6. Und angenommen, Du hättest ein Labor zur Verfügung - wie wäre Dein Vorgehen dann?

Aufgabe 45 (LC-MS Map Alignment). Wichtig für die Interpretation von verschiedenen LC-MS Experimenten ist das Alignment der Retentionszeiten, dass man gemeinhin als Warping bezeichnet. Dazu das Ihnen aus der Microarray-Normalisierung bekannte LOESS- (oder Lowess-) Verfahren - mit ein paar Anpassungen - verwendet.

1. Laden Sie die Daten von der Website¹ und lesen Sie sie in R. Die Daten sind durch Leerzeichen getrennt. Die beiden Spalten entsprechen den Retentionszeiten der einander entsprechenden Features in den beiden Datensätzen (x, y) .
2. Erstellen Sie den MA Plot, d.h. (x, y) wird ersetzt durch $((x + y), (y - x))$.
3. Führen Sie auf dem MA Plot die LOESS-Regression durch. Visualisieren Sie den Verlauf der Regressionsfunktion im MA plot.
4. Wenden Sie nun die so erhaltene Regressionsfunktion auf die ursprünglichen Daten (x, y) an.
5. Idealtypisch würden die Datenpunkte danach also entlang der Hauptdiagonale liegen. Um Overfitting zu vermeiden, verfahren Sie wie folgt: Berechnen Sie als Qualitätsmaß den mittleren absoluten Abstand der gewarpten Daten von der Hauptdiagonalen.
6. Optimieren Sie die Parameter der LOESS-Regression, wobei Sie die Regression nur auf ungeraden Datenzeilen durchführen und zum Ermitteln der Qualität der Warping-Funktion nur die geraden Datenzeilen verwenden.

Aufgabe 46 (Compomere). Bestimmen Sie mit dem Algorithmus aus der Vorlesung alle Compomere der Masse 13 über dem gewichteten Alphabet $\{3, 4, 5\}$. Erstellen Sie dazu die DP-Matrix und kennzeichnen Sie darin auch die Traceback-Pfade.

Aufgabe 47 (Isotopenverteilung). Jemand hat Fulleren C_{60} hergestellt aus Kohlenstoff mit folgendem Mengenverhältnis der Isotope: $^{12}C : 99\%$, $^{13}C : 1\%$. Berechnen Sie die Isotopenverteilung mit dem Algorithmus aus der Vorlesung. Tipp: Ein Tabellenkalkulationsprogramm kann hier hilfreich sein und Übersicht verschaffen.

Aufgabe 48 (Verständnisfragen). Im Skript zur Vorlesungen finden sich einige Fragen (*Exercise*). Bereiten Sie sie für die Übung vor, um ihr Verständnis der Materie zu überprüfen. Diese Aufgabe sind freiwillig und müssen nicht abgegeben werden.

¹http://lectures.molgen.mpg.de/Algorithmische_Bioinformatik_WS0809/