

# Projektbeschreibung 2

Freie Universität Berlin, SS 2005

Stefan Bentink · Jochen Jäger · Utz J. Pape · Claudio Lottaz · Rainer Spang

## Übungen zur Vorlesung Statistik für Bioinformatiker

Bearbeitungszeitraum: 14.06.2005 - 28.06.2005

- Arbeiten Sie in Zweier- oder Dreier-Gruppen. Schreiben Sie Ihre Namen und Matrikelnummern auf die erste Seite Ihres Projektberichts.
- Geben Sie Ihr Projekt in gedruckter Form (max 5000 Worte) zu Beginn der Vorlesung am 28.06.2005 ab. Senden Sie bitte zusätzlich den R-Code sowie den Bericht per Email an Ihren Übungsleiter.
- Zu jeder Aufgabe muss ein auf unserer R-Installation lauffähiger und ausführlich kommentierter Quellcode erstellt werden. Bioconductor Release 1.5 Packages dürfen benutzt werden. Starten Sie auf unseren Maschinen `/project/gene_expression/R-rel`, um Bioconductor-Packages zu benutzen.
- Beschreiben Sie in Ihrem Projektbericht die Benutzung Ihrer Funktionen und erörtern Sie nachvollziehbar Ihre Resultate.

## Allgemeine Hinweise zum Schreiben eines Berichtes

- **Einleitung:** Schreiben Sie einen Absatz zum Umfeld der Aufgabe und umreißen Sie Ihre Herangehensweise.
- **Methode:** Machen Sie detaillierte Angaben über das verwendete Material, die Daten und die benutzte Methode. Unter Umständen lohnt sich eine schematische Darstellung. Achten Sie darauf, dass der Leser Ihr Vorgehen nachvollziehen kann.
- **Resultate:** Veranschaulichen Sie die gefundenen Resultate. Auch hier sind Grafiken erwünscht.
- **Zusammenfassung/Diskussion:** Ziehen Sie Schlüsse aus den Beobachtungen.
- Illustrieren Sie Ihre Antworten mit vollständig beschrifteten Abbildungen und verwenden Sie unabhängig vom Fließtext verständliche Bildunterschriften.
- **Sprachstil:** Halten Sie den Text kurz und bündig. Versuchen Sie nicht, den Leser mit vielen Seiten zu beeindrucken. Ihre Kernpunkte erreichen den Leser am ehesten, wenn dieser nicht von unnötigen Informationen abgelenkt wird.

# 1 Einleitung

Mit Hilfe von cDNA-Microarrays oder DNA-Chips kann man die Aktivität (den Expressionswert) von mehreren tausend Genen eines Gewebes gleichzeitig messen. Die Technologie ermöglicht sehr detaillierte Einblicke in die molekularen Unterschiede von Geweben und Krankheiten. In diesem Projekt werden sie mit Microarraydaten (Genexpressionsdaten) von menschlichem Blutkrebs(Leukämie) analysieren. Der Beispieldatensatz stammt von einer Microarray-Studie (Genexpressionsprofil) von 128 Patienten mit B- oder T-Zell Leukämie. Die Microarrays der Firma Affymetrix, die hier verwendet wurden, messen die Expression von circa 12000 Genen, wobei die einzelnen Gene durch ein oder mehrere sogenannte "Probesets" repräsentiert werden. Daraus ergibt sich eine Datenmatrix mit 128 Spalten (für die Patienten) und mehr als 12000 Zeilen (für die einzelnen Probesets).

Die Anwendung von Genexpressionsprofilen zur Charakterisierung verschiedener Krebserkrankungen ist sehr neu und vielversprechend. Langfristig könnte man sich unter anderem eine Therapie von Krankheiten vorstellen, die entsprechend der Genexpressionsprofile individueller Patienten angepasst werden kann. Die Entwicklung und Etablierung statistischer Methoden zur Auswertung von Genexpressionsdaten ist ein spannender und interdisziplinärer Forschungsbereich. In dieser Übung sollen Sie ein Gefühl für den Umgang mit solchen Daten erhalten.

## Aufgabe 1 (Verteilung und Simulation von Genexpressionsdaten).

Genexpressionsdaten sind hochdimensionale Daten, die aus wirklichen Experimenten stammen. Über ihre statistische Verteilung kann man meist nur Vermutungen anstellen. Dennoch ist es von Zeit zu Zeit notwendig, dass man sich künstliche Expressionsdaten erzeugt (simuliert), um zum Beispiel neue Analyse-Methoden unter kontrollierten Bedingungen testen zu können. Zum Simulieren von Daten muss man aber Verteilungen und Parameter vorgeben.

- Besorgen Sie sich die Expressionsdaten. Sie finden sie auf unserer Webseite unter <http://lectures.molgen.mpg.de/statistik/material/p2/exprData.rdat>. Alternativ können Sie auch das Bioconductor Datenpaket ALL verwenden. Dieses beinhaltet die gleichen Daten.
- Machen sie sich mit der Objektklasse `exprSet` aus dem Bioconductor Paket `Biobase` vertraut. Die Daten, die Sie eben geladen haben sind von diesem Typ. Extrahieren sie daraus die eigentliche Expressionsmatrix mit dem Befehl `exprs(ALL)`.
- Ermitteln Sie mit Methoden, die Sie in der Vorlesung und in der Übung gelernt haben, welcher Verteilung die Zahlenwerte in der Expressionsmatrix unabhängig von Zeilen und Spalten am ehesten entsprechen. Vergleichen und diskutieren Sie mindestens drei verschiedene Verteilungsannahmen. Begründen Sie Ihre Entscheidung mit Hilfe grafischer Darstellungen.
- Erzeugen Sie eine Matrix gleicher Dimension wie die Expressionsmatrix und füllen Sie sie mit Zufallszahlen der gleichen empirischen Verteilung wie die Expressionsdaten. Man kann davon ausgehen, dass keine Ihrer oben vorgeschlagenen Verteilungsannahmen richtig ist. Trotzdem ist solch eine Datenmatrix zur Evaluation von Methoden sehr nützlich..
- Betrachtet man jedes Gen einzeln (jede Zeile der Expressionsmatrix), so erhält man jeweils eine Verteilung. Berechnen Sie Varianz und Mittelwert für jedes einzelne Gen. Schauen sie sich diese Parameter auch in Ihrer simulierten Matrix an. Vergleichen Sie die Verteilungen und beschreiben Sie Ihre Beobachtung.
- Erinnern Sie sich an das Übungsblatt zur linearen Algebra. Darin benötigten wir in etwa den folgenden Ausdruck:

$$\frac{\langle x, y \rangle}{\|x\| \|y\|}$$

Das Ergebnis ist eine Zahl zwischen -1 und 1, die dem Winkel zwischen zwei Vektoren proportional ist. Betrachtet man zwei Datenvektoren, dann kann man einen kleinen Winkel zwischen ihnen auch als

Abhängigkeit der beiden entsprechenden Zufallsvariablen interpretieren. In der Tat ist der empirische Korrelationskoeffizient wie folgt definiert:

$$r = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sqrt{\sum(x - \bar{x})^2} \sqrt{\sum(y - \bar{y})^2}}$$

Was ist der Unterschied zwischen  $r$  und dem Quotienten von Skalarprodukt und den beiden Normen? Implementieren Sie eine R-Funktion, die den empirischen Korrelationskoeffizienten zwischen zwei Vektoren berechnet.

- Die Korrelation zwischen den Expressionswerten von zwei Genen entspricht der Korrelation der zu den beiden Genen gehörenden Zeilenvektoren unserer Expressionsmatrix. Benutzen Sie die eben implementierte Funktion, um die Verteilung der paarweisen Korrelationen der Gene aus den Originaldaten mit denen Ihrer simulierten Daten zu vergleichen. Was beobachten Sie? Versuchen Sie sich an einer Interpretation. (**Tip:** Hier könnten Sie wieder an speichertechnische Grenzen Ihres Computers kommen. Denken Sie zum Beispiel daran, wie Prognosen in der Wahlberichterstattung erstellt werden.)

### Aufgabe 2 (Interpretation von genweisen Expressionswerten und Identifikation von Subgruppen).

- Jede Zeile unserer Expressionsmatrix hat einen Namen. Dieser Name entspricht einer ID, die der Hersteller der Microarrays, nämlich die Firma Affymetrix, vergeben hat. Sehen Sie sich die Expressionswerte der folgenden Probesets in allen Patienten an: `38355_at` und `38319_at`. Charakterisieren Sie die Verteilungen der Expressionswerte der einzelnen Gene so, wie Sie gelernt haben Verteilungen zu charakterisieren (auch grafisch). Vergleichen Sie die Ergebnisse auch mit der Expression anderer Gene. Was fällt Ihnen auf? Versuchen Sie mit Hilfe des Internets herauszufinden, welche Gene durch diese Probesets repräsentiert werden, zum Beispiel mit Hilfe der Webseite:

<http://apps1.niaid.nih.gov/david/upload1.asp>

Alternativ können Sie auch das Bioconductor Meta-Daten-Paket `hgu95av2` verwenden, welches die Annotationen enthält. Verwenden Sie die Informationen aus der Datenbank OMIM und Entrez-Gene (früher LocusLink) für eine biologische Interpretation. (**Tip:** Tumore werden nach dem Zelltyp klassifiziert aus dem Sie entstehen. Aus welchen Zelltypen sind die Tumore in unserem Datensatz entstanden?)

- Die beiden genannten Gene zeigen offensichtlich auffällige Verteilungen, die auch biologisch Sinn machen. Schlagen Sie vor, wie man solche Gene finden könnte:
  1. wenn Sie gar nichts über die Daten wissen,
  2. wenn Sie Gene, wie die beiden genannten kennen und noch weitere finden möchten.
- Suchen Sie jeweils 10 Gene (Probesets), die sich in den Patienten wie die beiden vorgegebenen Probesets verhalten.

### Aufgabe 3 (Differenzielle Genexpression zwischen verschiedenen Tumorentitäten).

Das Konzept der **differenziellen Genexpression** gab es schon bevor es Microarrays gab. Grob geht man dabei davon aus, dass biologisch unterschiedliche Gewebegruppen bestimmte Gene auch unterschiedlich stark exprimieren. In der Klinik nutzt man differenziell exprimierte Gene auch dazu Diagnosen zu erstellen. So könnte die Expression eines bestimmten Gens in Tumoren mit einer schlechten Prognose zum Beispiel stärker sein als in Tumoren mit einer guten Prognose. Dieses Marker-Gen wäre dann ein prognostischer Marker.

- Benutzen Sie den Befehl `pData(ALL)` um die phänotypischen Informationen zu extrahieren. Jede Zeile entspricht dabei einem Patienten und jede Spalte einer klinischen oder biologischen Variable. Die Spalte `BT` enthält für jedes Expressionsprofil die Information, ob es von einer B- oder einer T-Zelleukämie stammt. Prozessieren Sie die Informationen in der Spalte so, dass Sie einen Vektor mit T für T- und B für B-Zelleukämie erhalten. Machen Sie einen Barplot der Expression für das Probeset `1096_g_at`, wobei die B-Zell-Leukämien links und die T-Zell-Leukämien rechts stehen sollen. Benutzen Sie den Mittelwertsunterschied in beiden Gruppen als Maß für differenzielle Expression. Erstellen Sie eine Liste mit 1000 am stärksten differenziell exprimierten Genen.

- Beschreiben Sie in eigenen Worten, was man mit der folgenden Formel erhält, wenn  $X$  die Expression eines Gens in einer Gruppe von  $m$  Patienten und  $Y$  die Expression in einer anderen Gruppe von  $n$  Patienten ist, mit  $S^2$  als Varianz der Zufallsvariablen  $X$  und  $\tilde{S}^2$  als Varianz der Zufallsvariable  $Y$ .  $\bar{X}$  ist der Expressions-Mittelwert.

$$T = \sqrt{\frac{m * n * (m + n - 2)}{m + n}} * \frac{\bar{Y} - \bar{X}}{\sqrt{(m - 1)S^2 + (n - 1)\tilde{S}^2}}$$

- Implementieren Sie eine Funktion in R die den Ausdruck  $T$  berechnet. Bestimmen Sie mit dieser Funktion die Gene mit den 1000 absolut höchsten T-Werten. Sie haben nun eine zweite Liste mit differentiell exprimierten Genen. Vergleichen Sie diese Liste mit der vorherigen Liste. Gibt es Unterschiede? Wenn ja, erklären Sie diese. Sehen Sie sich dazu auch die Expression von solchen Genen an, die in der einen Liste sind aber nicht in der anderen. Das können Sie zum Beispiel mit dem eben erzeugten Barplot machen.

#### Aufgabe 4 (Häufung bestimmter funktioneller Gruppen und Pathways in den am stärksten differentiell exprimierten Genen).

Bei der Analyse von Genexpressionsdaten ist man häufig daran interessiert welche Funktionen die differentiell exprimierten Gene haben. Spielen viele dieser Gene eine Rolle in der Zellteilung so könnte man zum Beispiel die Hypothese formulieren, dass der molekulare Unterschied zwischen den verglichenen Gruppen etwas mit der Zellteilung zu tun hat. Für solche Analysen benutzt man häufig die sogenannte Gene Ontology (<http://www.geneontology.org>). Hierbei handelt es sich um eine Datenbank, die eine Terminologie zur Beschreibung von Genfunktionen zur Verfügung stellt. Mit dieser vereinheitlichten Beschreibung ist es leicht möglich Gene zu funktionellen Gruppen zusammenzufassen. Ausgehend von einer Liste differentiell exprimierter Gene ergibt sich die folgende Situation:

	differentiell exprimiert	nicht differentiell exprimiert
funktionelle Gruppe Zellteilung	Anzahl?	Anzahl?
andere funktionelle Gruppen	Anzahl?	Anzahl?

In dieser **Kontingenztafel** wird gegenübergestellt wieviele der differentiell exprimierten Gene zu einer bestimmten funktionellen Gruppe gehören und wieviele nicht. Sie beschreibt einen möglichen Ausgang eines Zufallsexperiments. Die Wahrscheinlichkeit für diesen Ausgang lässt sich aus einem Urnenmodell ableiten und folgt einer hypergeometrischen Verteilung.

- Erzeugen Sie eine Liste mit 500 differentiell exprimierten Genen. Verwenden Sie dabei den Vergleich aus der vorherigen Aufgabe. Das Mass für differentielle Expression bleibt Ihnen überlassen. Das R-Objekt <http://lectures.molgen.mpg.de/statistik/material/p2/goGroups.rdat> enthält eine Liste mit funktionellen Gruppen. Dabei entspricht der Name jedes Listenelements der Bezeichnung für die jeweilige funktionelle Gruppe. Die Listenelemente selbst enthalten jeweils einen Vektor vom Typ Character, der die Namen der einzelnen Probesets enthält, die den funktionellen Gruppen zugeordnet werden. Schreiben Sie eine R-Funktion, die für eine gegebene funktionelle Gruppen die Kontingenztafel erzeugt.
- Berechnen Sie nach dem La-Place Modell die Wahrscheinlichkeit, dass gleich viele oder mehr Gene einer zufälligen Liste der gleichen Länge in diese funktionelle Gruppe fallen.
- Machen Sie sich mit der R-Funktion `fisher.test()` vertraut. Was erwartet diese Funktion als Eingabe? Wie sieht die Ausgabe aus? Berechnen Sie obige Wahrscheinlichkeit mit dieser R-Funktion.
- Manche Gene werden durch mehrere Probesets repräsentiert und tauchen daher mehrmals in der Liste differentiell exprimierter Gene auf. Welchen Einfluss hat das auf Ihre eben beschriebene Analyse. Wie können Sie dafür korrigieren? Das R-Objekt <http://lectures.molgen.mpg.de/statistik/material/p2/lokusIds.rdat> enthält eine Liste mit Gen-Ids für jede Zeile Ihrer Datenmatrix. Damit können Sie Gene finden, die mehrfach in Ihren Listen auftauchen. Korrigieren Sie Ihre Analyse entsprechend.