Softwarewerkzeuge der Bioinformatik

Prof. Dr. Volkhard Helms — Saarland University
PD Dr. Michael Hutter, Kerstin Reuter, Daria Gaidar — Department of Computational Biology
Wintersemester 2017/2018

Tutorial 8

4. Januar 2018

Tutoren: Lea Eckhart, Markus Dillmann

Gene Expression

Im heutigen Tutorial geht es um die Analyse von Genexpressionen aus Microarray-Experimenten. Mit Hilfe des "*MicroArray Genome Imaging & Clustering Tool*" (MAGIC Tool) werden Sie mit einfachen, nachgemachten Bildern den kompletten Prozess von der Bild- bis zur Clusteranalyse durchführen. Die roten Bilder beschreiben die zeitabhängige Expression von neun hypothetischen Genen an fünf aufeinanderfolgenden Zeitpunkten, die grünen dienen als Kontrolle und geben jeweils die Expression zum ersten Zeitpunkt auf den verschieden reagierenden Microarrays wieder. Weitere Infos finden Sie unter http://www.bio.davidson.edu/projects/magic/magic.html

Exercise 8.1: Vorbereitung

(a) Bitte laden Sie Magic Tool 2.1 herunter. Und starten Sie es mit einem Klick auf `MagicTool.jar`. http://www.bio.davidson.edu/projects/magic/magic.html

(b) Alle für dieses Tutorial benötigten Dateien sind diesmal online auf unserer Kurs Homepage https://www-cbi.cs.uni-saarland.de/teaching/...

(c) Öffnen Sie dieses neue Projekt Über den Menübefehl: `Project > Load`. Nun können Sie mit dem Einlesen der Bilder die Analyse beginnen.

Exercise 8.2: Einlesen der Microarray-Bilder

(a) Laden Sie das erste rote Bild über die Menüauswahl (oder das äquivalent Tastaturkürzel `Ctrl + R`):

```
Build Expression File > Load Image Pair... > Red
```

Der Dateiname lautet "`red1.gif`" (zum Laden müssen Sie den Dateifilter auf "`*.gif`" umschalten). Laden Sie analog das grüne Bild (`green1.gif`) (`Ctrl + G`).

(b) Sie benötigen eine Zuordnung der Gene zu den entprechenden Spots. Öffnen Sie das Fenster zum Laden der Genliste mit: (`Ctrl + X`)

```
Build Expression File > Load Gene List...
```

Die Datei `myGeneList.txt` enthält die notwendigen Annotationen.

(c) Als nächstes definieren Sie das Gitter, um die Spots auf dem Array zu kennzeichen. Öffnen Sie den Dialog mit: (`Ctrl + A`)

```
Build Expression File
> Addressing/Gridding
> Create/Edit Grid
```

Lesen und bestätigen Sie die Warnung. Achten Sie darauf, dass die Spotreihenfolge im folgenden Dialog korrekt angegeben ist. Die Spots sind so angeordnet, dass Nummer 1 links oben ist, Nummer 2 und Nummer 3 darunter. Nummer 4 ist der oberste Spot in der nächsten Reihe (keine default Einstellung, unten im *Grid Setup* auf *vertically* stellen).

Um das Gitter zu definieren, klicken Sie `Set Top Left Spot` und klicken Sie im Bild auf den linken oberen Spot. Wiederholen Sie dies für den oberen rechten Spot und einen Spot der untersten Reihe. Nachdem Sie die Anzahl der Zeilen und Spalten eingetragen haben, erscheint mit einem Klick auf `Update` ein Gitter über dem Bild der Spots. Das Gitter passt wahrscheinlich nicht ganz. Ziehen Sie es mit der Maus an den weissen Punkten, sodass die neun Spots möglichst in den Mitten der einzelnen Zellen liegen. Wenn Sie damit zufrieden sind, bestätigen Sie mit `Done!`. Wenn die Meldung kommt, dass Sie neun Gene "*gridded and named*" haben, hat dieser Schritt funktioniert. Speichern Sie das Gitter ab, Sie werden es noch brauchen.

(d) Jetzt können Sie über `Build Expression File > Segmentation` (`Ctrl + S`) das Einlesen der Helligkeitswerte starten. Lassen Sie die `Segmentation Method` auf `Fixed Circle`, erhöhen Sie jedoch den Radius soweit wie möglich, so dass die roten Kreise immer noch innerhalb der Spots liegen (`Update Data` nicht vergessen). Weiter unten können Sie mit dem `Next`-Button durch die Spots schalten und überprüfen, ob der Kreis im roten und im grünen Bild korrekt über den Spots liegt. Wenn Sie zufrieden sind und keiner der Spots als *"Automatically Flagged"* gekennzeichnet ist, erstellen Sie das Expressionsprofil (`Create Expression File` weiter unten in der linken Leiste). Im folgenden Dialog geben Sie als `Expression Filename` `eins_t1` ein und als `Column Name` `t1`, (die ersten Bilder entsprechen dem ersten Zeitpunkt). Bestätigen Sie mit `OK`, das Fenster geht zu.

(e) Unter `Expression > Working Expression File` sehen Sie, dass das gerade erstellte Expressionsprofil ausgewählt ist. Über `Expression > View Data` können Sie die extrahierten Daten ansehen. Alle Werte sollten nahe bei Eins liegen. Schliessen Sie dieses Fenster mit der Liste wieder (Knopf rechts oben).

(f) Für den nähsten Zeitpunkt laden Sie `red2.gif` und `green2.gif` wie oben beschrieben. Die Genliste ist die selbe wie vorher und muss nicht erneut geladen werden. Laden Sie dann das vorher gespeicherte Gitter ("`Build Expression File > Addressing/Gridding > Load Saved Grid`") und überprüfen Sie das Gitter unter
`Build Expression File > Addressing/Gridding > Create/Edit Grid`.

(g) Rufen Sie dann `Build Expression File > Segmentation` auf und erhöhen Sie wieder den Radius. Einmal durch die Spots klicken und `Create Expression File`. Geben Sie als Dateinamen `eins_t2` und als `Column name` `t2` ein. Wählen Sie `Append to File` und wählen Sie aus dem Popup-Menu das Group file `eins_t1/eins_t1.exp` aus. `OK`.

(h) Laden Sie nun die Bilder `red3.gif` und `green3.gif` für den dritten Zeitpunkt, dann das Gitter, etc. Nach `Create Expression File` geben Sie den Dateinnamen `eins_t3` und die Spaltenbezeichnung `t3` ein. Hängen Sie diese Daten an `eins_t2/eins_t2.exp` an. Verfahren Sie analog für den 4. und 5. Zeitpunkt. Am Ende sollten in `eins_t5/eins_t5.exp` fünf Spalten mit Expressionsverhältnissen stehen (`Expression > View Data`).

Exercise 8.3: Datenanalyse

Achten Sie im folgenden Teil darauf, dass unter `Expression > Working Expression File` jeweils das richtige Expressionsprofil ausgewählt ist. Wird eine neue Datei angelegt, so wird diese automatisch ausgewählt.

(a) Rufen Sie `Expression > Explore` (`Ctrl + E`) auf. Klicken Sie auf `Plot Selected Group`. Sie können in diesem Plot gut erkennen, welche Gene wann Überexprimiert sind, doch unterexprimierte Gene sind kaum zu erkennen. Deshalb werden wir die Expressionskurven von direkten Intensitätsverhältnissen in log-Ratios transformieren.

(b) Rufen Sie dazu `Expression > Manipulate Data > Transform` aus und wählen Sie $\log_b$ mit $b = 2$. Akzeptieren Sie den vorgeschlagenen Dateinamen (er hat `_tlog2` angehängt).

(c) Rufen Sie erneut `Explore` auf und plotten Sie den ganzen Datensatz. Nun sind auch unterexprimierte Gene gut zu erkennen. Wenn Sie einen der Punkte im Plot anklicken, wird der zugehörige Verlauf rot hervorgehoben. Bei gedrückter Shift-Taste können Sie mehrere Verläufe markieren; oberhalb des Plots kann ein Panel mit Beschreibungen aufgezogen werden.

Exercise 8.4: Clustering

Bevor die Expressionprofile der neun hypothetischen Gene nach ähnlichem Verlauf sortiert (geclustert) werden können, muss zuerst die *dissimilarity* zwischen den Kurven bestimmt werden.

(a) Rufen Sie dazu `Expression > Dissimilarities > Compute` (`Ctrl+D`) auf. Lassen Sie die Methode auf dem Default $1 - \text{correlation}$. Als Ausgabedatei sollte `eins_t5_tlog2.dis` vorgeschlagen sein. Mit den "dissimilarities"können die Gene nun mit verschiedenen Clustertypen aufgeteilt werden.

(b) Unter `Cluster > Compute` wählen Sie nun das gerade erstellte dissimilarity-file und als Methode weiter unten *Hierarchical Clustering*. Dieser Typ wird im `Cluster file` mit einem angehängten "h" gekennzeichnet. Ist alles eingestellt, wird die Berechnung mit `OK` gestartet.

(c) Unter `Cluster > Display` stehen nun verschiedene Darstellungen zur Verfügung. Wählen Sie den gerade berechneten Cluster und stellen Sie ihn als erstes als `Metric Tree` dar. Die Skala oberhalb des Graphs gibt die Dissimilarity an. Per Mausklick auf die schwarzen Vierecke im Baum können Sie einen Subgraphen auswählen, dessen Expressionsprofile mit `Plot Selected Nodes` angezeigt werden kann. Überprüfen Sie, ob Sie die Gene genauso geclustert hätten. Schliessen Sie die Plots wieder. Für grössere Datenmengen eignet sich auch die Darstellung als `Exploding Tree`.

(d) Stellen Sie nun die Cluster als `Tree/Table` dar (mit 20 als `Line Height`, `Update` nicht vergessen). Versuchen Sie nun, mit den Reglern den Dynamikbereich so einzustellen, dass die Unterschiede zwischen den Gruppen ORFA/ORFC und ORFB/ORFH *augenscheinlich* werden. (die Grenze zu einer suggestiven, verzerrenden Manipulation der Daten ist dabei nicht weit weg).

(e) Erstellen Sie nun einen `QT-Cluster` mit maximal 9 Clustern und der minimalen Grösse von 1 und vergleichen Sie in der `Exploding Tree`-Darstellung die beiden berechneten Cluster.

Lassen Sie sich von einzelnen Clustern die Verläufe anzeigen. Was bemerken Sie? Sind die Zuordnungen vergleichbar? Gibt es Gene, die schwerer einzuordnen sind?

Have fun!