Bioinformatik in der Arzneistoffforschung
im SoSe 2023

Blatt 2

Prof. Dr. Andreas Dominik

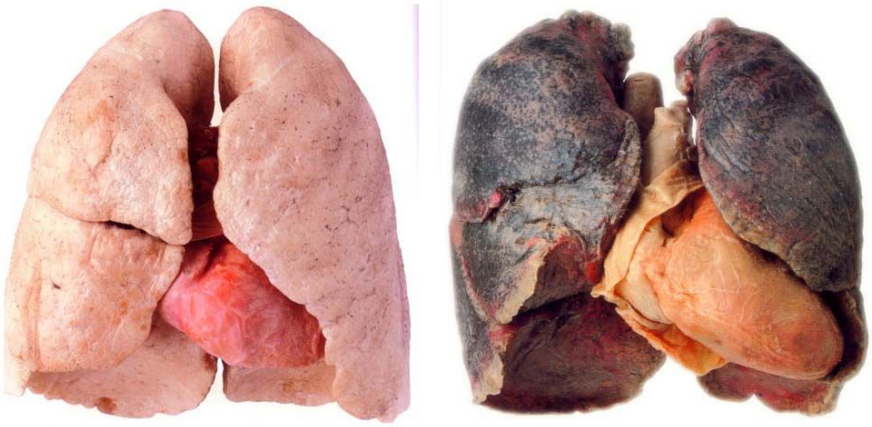
Spielregeln

- Die Aufgaben sollen im Praktikum und zuhause einzeln bearbeitet werden.
- Anwesenheit im Praktikum und Abgabe der Lösungen ist Pflicht für die Teilnehmer, des Kurs mit 6 CrP als WP-Modul im Bachelor.
  Für die Teilnahme am Methodenseminar mit nur 3 CrP (Masterstudiengang Bioinformatik und Systembiologie) ist die Teilnahme am Praktikum freiwillig.
- Die Lösung muss in Moodle abgegeben werden. Bitte als *ordentliches* tar-Archiv; d.h. als eine einzige Datei mit dem Namen `Nachnahme.Vorname_Blatt_Nummer.tar.gz`.
  Das Archiv soll nur ein Unterverzeichnis (`Nachnahme.Vorname-Blatt-Nummer/`) enthalten in dem alle Dateien und ggf. weitere Unterverzeichnisse zu finden sind.
- Die Übung wird in den Praktika bis zum **15. Juni 2023** besprochen und muss an diesem Tag bis spätestens 23:59h in Moodle hoch geladen werden. Da Moodle dann dicht macht können die Aufgaben weder nachträglich noch als e-Mail abgegeben werden.

Was macht das Rauchen mit unserer Lunge? Analyse der differenziellen Genexpression (20)

1 Daten

Die Molekularbiologen können heute unser Gewebe sehr genau untersuchen, um Krankheiten präzise diagnostizieren zu können. Bei diesem Beispiel geht es um die Veränderungen, die in der Lunge von Rauchern vor sich gehen. Diese Unterschiede sind sicher sehr groß, wie man schon am Aussehen einer Lunge sehen kann (es kann sicher jeder leicht erkennen, welche der beiden abgebildeten Lungen von einem Raucher stammt):



Eine geschädigte Lunge beschert dem Besitzer häufig die Krankheit COPD (chronic obstructive pulmonary disease). Das äußert sich vielleicht nur als Raucherhusten, ist aber tatsächlich eine Schädigung der Lunge, die irreversibel ist (also auch nicht mehr abheilt, wenn man aufhört zu Rauchen).

Das Ausmaß der Schädigung zu erkennen und vielleicht zu lindern wäre deshalb eine gute Sache - wenn sich Molekularbiologie und Bioinformatik zusammentun …kein Problem!

In jeder unserer Körperzellen sind ja alle unsere Gene gespeichert. Damit die Zellen unterschiedliche Funktionen ausführen können, schaltet jede Zelle bestimmte Gene ein oder aus. Dies ist aber nicht ein binäres Ein/Aus sondern kann ein gleitender Übergang sein (von aus über ein bisschen ein, etwas mehr ein bis ganz stark ein). Die wissenschaftliche Bezeichnung dafür ist *wenig exprimiert* oder *hoch exprimiert*.

Eine Analyse der Daten könnte ganz interessant sein: Gene, die durch das Rauchen eingeschaltet werden, können für die negativen Folgen verantwortliche sein. Wenn es möglich ist, ein Medikament zu entwickeln, dass diese Gene wieder ausschaltet, könnte man verhindern, dass durch das Rauchen Schäden entstehen.

Tatsächlich wurde bei dieser Studie die *Expression* fast aller unserer Gene gleichzeitig gemessen. Dies erzeugt ca. 54.000 Messwerte für jeden der untersuchten 22 Patienten im Datensatz. Insgesamt ergibt das eine Tabelle mit 22 Spalten und ca. 54.000 Zeilen.

Leider sind die Biologen beim Datamanagement nicht sehr sorgfältig:

- Es kann sein, dass einzelne Werte fehlen (d.h. eine Ziele hat weniger als 22 Werte). Diese Zeilen müssen dann ignoriert werden.
- Die Namen der Records sind nicht aussagekräftig. Da stehen mehr oder weniger willkürliche IDs, die vom Hersteller des Messgeräts vergeben werden. Man benötigt eine extra Tabelle, in der diese IDs mit den Namen und Beschreibungen der Gene verknüpft werden. Diese benötigen Sie aber erst am Ende der Analyse um die gefundenen Gene zu benennen.
- Es gibt tatsächlich nicht für alle IDs auch Gene in der Tabelle (diese können Sie bei der Auswertung ignorieren.
- Die Einteilung in Raucher/Nichtraucher steht nochmal in einer eigenen Datei.

Laden Sie den Datensatz Raucherlunge aus Moodle; er enthält 3 Dateien:

`GSE4498_series_data.txt`: Tabelle mit der Aktivität der Gene in Gewebeproben aus der Lunge von Rauchern und Nichtrauchern

HG-U133-GPL570-39741.txt: Beschreibung der Gene nach ID

metadata.txt: Klassifizierung der Proben nach Raucher oder Nichtraucher.

2 Differenzielle Genexpressionsanalyse

Die Aufgabe besteht darin, Unterschiede in der Genexpression zwischen Rauchern und Nichtrauchern zu finden; also: welche Gene sind in der Lunge von Rauchern stärker *eingeschaltet* als bei Nichtrauchern? erzeugen Sie eine Liste der Gene (nach Möglichkeit nicht die komischen IDs, sondern die Namen und Beschreibungen der Gene). Wenn Sie Bioinformatikerin oder Bioinformatiker sind können Sie sich vielleicht auch noch Gedanken darüber machen, was das für den Menschen bedeutet?

2.1 Ganz billige Lösung (z.B. in R)

Ein erster vereinfachter Lösungsansatz könnte sein, einfach für jedes Gen die Mittelwerte aller Raucher und aller Nichtraucher zu berechnen und dann ins Verhältnis zu setzen.

Dies können Sie mit R machen - das eignet sich perfekt dafür. Mit etwas Geschick beötigen Sie nur 5 Zeilen Code für alles.

Da in R alle Vektoroperationen elementweise (d.h. als Broadcast) ausgeführt werden, lässt sich das in R in wenigen Zeilen umsetzen. Da R eine interpretierte Sprache ist, sollten Sie wie in Python streng darauf achten, keine Schleifen zu verwenden (oder nur dort, wo es nicht weh tut).

Das Lesen der Daten könnte in R z.B. so aussehen:

```
x <- read.table("GSE4498_series_data.txt", header=TRUE, row.names = 1,
                stringsAsFactors = F)
classes <- read.table("klassen.txt", header = F, stringsAsFactors = F)
smokers <- classes[,2] == "smoker"
```

`read.table()` liefert einen Dataframe.

`smokers` ist danach ein Vektor mit boolschen Werten, der verwendet werden kann, um die Spalten der Raucher bzw. Nichtraucher zu addressieren.

Da Dataframes sowohl als Array (`x[2,5]`) als auch als Liste aus Spalten (`x[[1]]` oder `x$GSM0101095`) als auch mit einem boolschen Vektor (`x[2,smokers]`, `x[smokers]`) addressiert werden können, ist es leicht mit diesem Vektor `smokers` jeweils die richtige Auswahl zu treffen.

2.2 Bessere Lösung mit etwas Statistik (in R und KNIME)

Das mit den Mittelwerten ist natürlich nur die Lösung *für Dummies* - echte Wissenschaftler erkennen sofort, dass dieses Experiement ein ideales Beispiel für einen Hypothesentest darstellt:

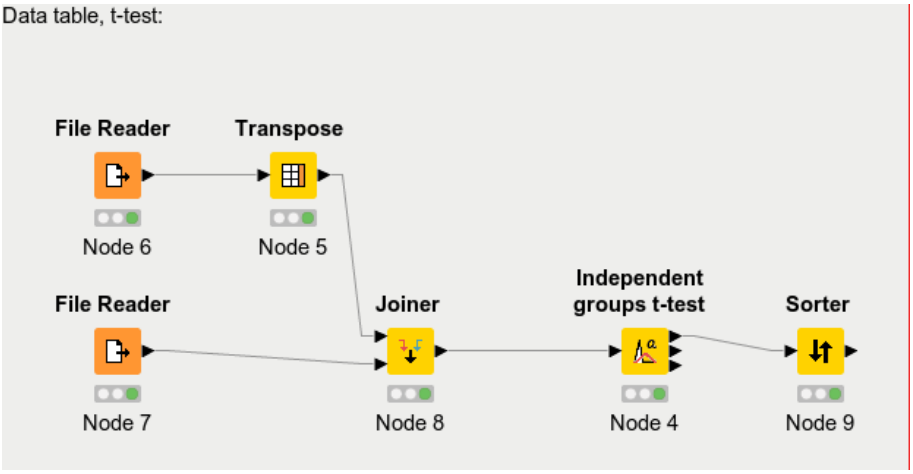
Es gibt 2 unabhängige Stichproben (ein paar gemessene Raucher und ein paar gemessene Nichtraucher) die jeweils (vielleicht näherungsweise - bitte wenigstens visuell prüfen!) standarderteilt streuen und mit etwas Glück auch eine ähnliche Varianz haben. Der sog. 2-Stichproben-t-Test ist das geeignete Mittel, um zu prüfen, ob die tatsächlichen Gesamtheiten (alle Raucher und alle Nichtraucher) aus denen die kleinen Stichproben stammen, sich unterscheiden (d.h. denselben Mittelwert haben). Der t-Test liefert und dazu den sog. P-Wert, der angibt mit welcher Wahrscheinlichkeit dieses Messergebnis gemessen wird, wenn die Mittelwerte tatsächlich gleich sind. Der P-Wert sollte deshalb möglichst klein sein (wissen Sie natürlich).

Zur Auswahl statistisch signifikant unterschiedlich exprimierter Gene müssen Sie deshalb nur den t-Test für jedes Gen (auf jede Zeile) anwenden und dann nach dem P-Wert filtern. Wir schauen uns alle Gene an, bei denen $P \leq 0,01$ ist (d.h. es bleibt 1% Wahrscheinlichkeit, dass die gemessenen Unterschiede nur Zufall sind und tatsächlich gar kein Unterschied besteht).

Das geht auch wieder leicht mit R, die magische Zeile, die alle P-Werte berechnet, lautet (wenn Sie nicht herausfinden, was das soll, dann können Sie es auch mit einer Schleife machen; spätestens im Praktikum werde ich versuchen Ihnen diese merkwüdigen R-Konstrukte zu erklären. Glauben Sie mir: wenn man's kann macht es Spaß!):

```
p <- as.data.frame(apply(x, 1, function(row){
                          t.test(row[smokers], row[!smokers])$p.value
                          }))
```

Oder Sie spielen mit KNIME, das mit der Datenmenge auch ganz gut umgehen kann. Es gibt für alle Teilaufgaben ganz gut geeignete Knoten. Der erste Teil des Workflows könnte z.B. so ähnlich aussehen:



3 Clustering

Laden Sie die Messdaten (GSE4498_series_data.txt) in knime (oder irgendetwas Anderes), transponieren Sie ggf. die Tabelle (so dass Patienten (GSE***) die Muster und Gene (20235_s_at) die Attribute sind und fügen Sie Namen und Beschreibung der Gene hinzu.

3.1 Ihre Aufgabe

- Versuchen Sie nun die Patienten in 2 Klassen zu clustern - es sollten dabei die Raucher von den Nichtrauchern unterschieden werden.
- Verwenden Sie verschiedene Algorithmen zum Clustern - welche machen hier Sinn?.
- Versuchen Sie herauszufinden, welche Gene die kranken Lungen von den gesunden Lungen unterscheiden (alle Raucherlungen sind krank).

Das Clustering erfolgt unüberwacht: Die Datei klassen.txt soll nur zur Kontrolle der Ergebnisses verwendet werden.

Achtung: Das dauert - je nach Ihrem Vorgehen - sehr lange! Zum Spielen empfiehlt es sich, einen verkleinerten Datensatz zu bauen.

Sie können jede beliebige Plattform oder Programmiersprache verwenden.

3.2 Statistische Attributselektion

Reduzieren Sie die Dimensionen von 50000 auf 500, indem Sie nur Attribute verwenden, deren Varianz über einem Limit liegt und

- Wiederholen Sie das Clustering (jetzt geht auch HCA).
- Vergleichen Sie die Ergebnisse mit und ohne Normalisierung (Achtung: dabei kann man auch Fehler machen!).

3.3 Gezielte Attributselektion

Wählen Sie ca. 200 Gene aus, deren Expression sich am stärksten zwischen Rauchern und Nichtrauchern unterscheidet (entweder ist die Expression höher oder niedriger). Dazu können Sie das Ergebnis der differenziellen Expressionsalanyse verweden!

Mit dieser reduzierten Zahl an Attributen können Sie fast beliebige Clusteringmethoden verwenden - zumindest technisch müsste alles funktionieren.

Versuchen Sie verschiedene Arten der Skalierung/Normalisierung.

- Visualisieren Sie die Daten. Lassen sich Raucher und Nichtraucher trennen?
- Ist jetzt eine Trennung der Raucher von den Nichtrauchern zu sehen?
- Wie unterscheiden sich die Ergebnisse?
- Versuchen Sie eine Visualisierung (es sind ja nur 22 Patienten und 200 Gene).
- Fragen: Welche Gene sind besonders stark unterschiedlich exprimiert? Können Sie herausfinden, was diese Gene bewirken oder warum diese bei Rauchern so stark/wenig aktiv sind?