Metabolomics

"Quantitative measurement of the dynamic multiparametric metabolic response of living systems to pathophysiological stimuli or genetic modification." (Jeremy Nicholson)

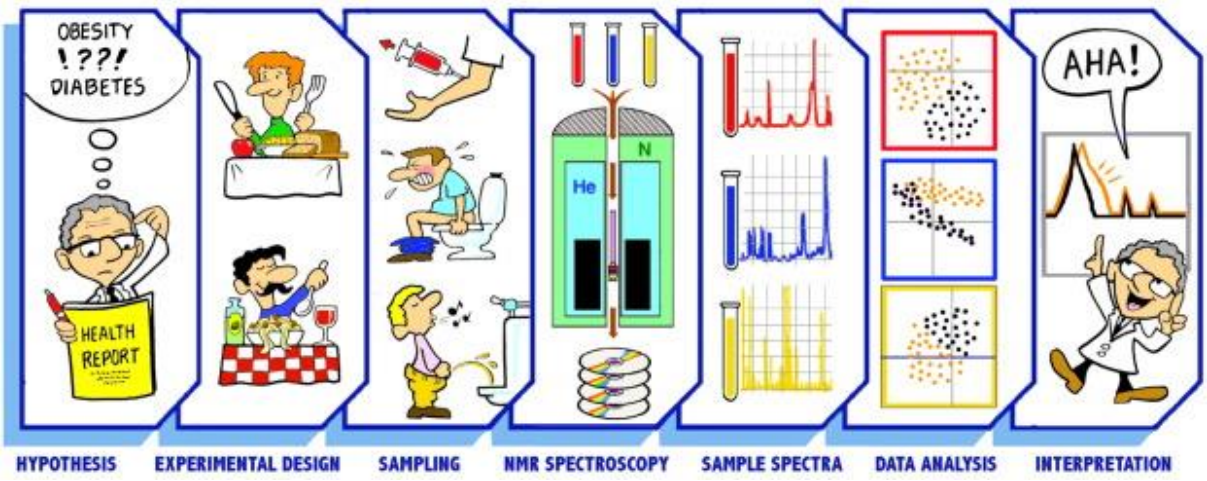
Einleitung

Metabolomics oder auch Metabonomics ist neben Genomics, Transcriptomics und Proteomics ein weiterer Bereich der "omics" Forschungsbereiche. Dieses Forschungsfeld befasst sich mit kleinen metabolischen Molekülen in biologischen Systemen. Die Begriffe Metabonomics und Metabolomics entwickelten sich zur selben Zeit in verschiedenen Arbeitsbereichen wie der Biochemie, Zoologie, Botanik oder der Mikrobiologie. Beide Bezeichnungen beinhalten die Messung von Metaboliten in Organismen, Geweben, Sekreten oder Biofluiden etc..

Metabonomics befasst sich mit biologischen Systemen (z.B. von Geweben) oder auch Biofluiden, während sich Metabolomics mit einfachen Zellsystemen oder intrazellulären Vorgängen beschäftigt. Meist werden diese Begriffe jedoch allgemein für die Messung, Identifizierung und Quantifizierung von Metaboliten verwendet. Im Folgenden werden diese Bezeichnungen als Synonym behandelt.

Sowohl die analytischen Methoden, welche für Transcriptomics oder auch Proteomics entwickelt wurden, als auch die Entwicklung von umfangreichen Datenbanken, ist für die Aufklärung der Metabolite unumgänglich. Ein Ziel der Metabonomicsforschung ist es, einen Überblick über den metabolischen Status eines Organismus zu bekommen und Anzeichen für Veränderungen in diesem feststellen zu können. Dies ist besonders im Hinblick auf einen pathologischen Status interessant. Eventuell können so einzelne Krankheitsstadien und deren Abfolgen am Metabolom festgestellt werden.

In diesem Praktikum werden wir eine kurze Einführung in die wichtigsten Aspekte bezüglich der Extraktion und Probenvorbereitung am Besiepiel Kaffee und der standardisierten Analyse mittels NMR (Nuclear magnetic resonance spectroscopy) sowie der statistischen Datenauswertung geben.



Savorani et al, A primer to nutritional metabolomics by NMR spectroscopy and chemometrics, In Food Research International, Volume 54, Issue 1, 2013

Theorie zur statistischen Datenauswertung

Metabolomicsstudien generieren immer eine extrem große Datenmenge. Auch wenn die Anzahl der Proben überschaubar bleibt (im Idealfall werden allerdings sehr viele Proben gemessen) enthält jedes aufgenommen 1D-Spektrum eine Fülle von Informationen. Um dieser Datenmasse Herr zu werden, muss man sich statistischer Methoden bedienen. Der grundlegende Ablauf ist immer ähnlich: die Spektren werden in Abschnitte zerlegt, sogenannte Buckets. Diese können automatisch generiert werden, sodass das Spektrum in Abschnitte einer festen Breite unterteilt wird oder manuell, wodurch gewährleistet werden kann, dass jeder Peak von einem Bucket umfasst und nicht auf mehrere verteilt wird. Letzteres stellt allerdings schon einen Eingriff in die Datenmatrix dar. Die Integralwerte der Buckets sind direkt proportionale zur Konzentration der Substanzen, deren Peaks erfasst wurden. Mit diesen Werten wird eine Datenmatrix erzeugt, in der jedem Spektrum für jedes Bucket ein Wert zugeordnet wird. Anhand dieser Matrix wird die statistische Analyse durchgeführt. Selbst wenn recht breite Buckets gewählt werden ist deren Zahl in den meisten Fällen trotzdem deutlich größer als die der Proben. Die meisten statistischen Methoden basieren jedoch auf der Annahme, dass die Zahl der Proben die der Variablen übersteigt und dass die Variablen normalverteilt sind. Beides ist in der Regel jedoch nur erfüllt, wenn sehr große Studien durchgeführt werden.

Der erste Schritt der statistischen Analyse ist häufig die Normalisierung und/oder Skalierung der Daten. Durch die Wahl der Funktion kann der Schwerpunkt der Studie gesetzt werden, wird jedoch mit einer unpassenden Methode skaliert kann dies das Ergebnis negativ beeinflussen.

Die Normierung auf Gesamtintensität soll einen Vergleich der Proben miteinander ermöglichen, da so der Verdünnungsfaktor eliminiert werden kann.

$$\tilde{x}_{ij} = \frac{x_{ij}}{\sum_{j=1}^{J} x_{ij}}$$

Unterschiedliche Gesamtkonzentrationen der verschiedenen Spektren können durch Abweichungen bei der Probenvorbereitung entstehen, zum Beispiel, wenn unterschiedliche Probenmengen eingewogen werden oder die Extraktion mal mehr und mal weniger effizient verläuft. Gerade bei Urinproben kann jedoch ein von der Probenpräparation unabhängiger Verdünnungseffekt auftreten, je nachdem, wie viel der Patient vor der Probennahme getrunken hatte. Der Nachteil der Normierung ist allerdings, dass Proben, die durch wenige sehr hoch konzentrierte Substanzen dominiert werden, zum Beispiel Zucker, letztlich auf deren Intensität normiert werden, man also impliziert, dass in jeder Probe die gleiche Gesamtmenge an Substanzen vorhanden sein muss. Es gibt noch andere Normierungsmethoden, wie zum Beispiel die Normierung auf den Median. Bei dieser Methode wird der median jeder Variablen bestimmt und vom eigentlichen Wert abgezogen. Das bedeutet, dass in der Datenmatrix für jeden Wert nur noch der Abstand zum Median angegeben wird. Andere Methoden setzen voraus, dass entweder eine Referenzprobe oder eine Referenzsubstanz gemessen wurde.

Zudem werden die Daten im Allgemeinen zentriert, das heißt, für jeden Wert wird der Abstand zum Mittelwert berechnet:

$$\tilde{x}_{ij} = x_{ij} - \bar{x}_i$$

Die normierten Daten werden dann skaliert. Die beiden am häufigsten benutzten Methoden sind Paretoskalierung und Unitvariance-Skalierung (in Metaboanalyst auto scaling genannt). In beiden

Fällen wird für jede Variable ein eigener Skalierungsfaktor berechnet, der die Werte der einzelnen Variablen (zum Beispiel Konzentrationen) Wertunterschiede relativ zum Skalierungsfaktor transformiert und damit vergleichbarer macht. Ohne Skalierung wird die statistische Auswertung sonst unter Umständen durch wenige sehr hoch konzentrierte, aber unter Umständen eigentlich nicht relevante Substanzen dominiert. Kleine Werte werden dann oft in der Relation verstärkt, wodurch der für kleine Konzentrationen ohnehin schon große Fehler nochmals vergrößert wird.

Unitvariance-Skalierung verwendet die Standardabweichung als Skalierungsfaktor:

$$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_i}{\sigma_i}$$

Nach der Skalierung haben alle Variablen eine Standardabweichung von 1, das heißt die nachfolgenden Datenanalysen beruhen auf der Korrelation und nicht mehr der Kovarianz. Nach der Skalierung sind die Daten dimensionslos. Durch diese Skalierungsmethode sind alle Variablen gleich wichtig für die folgenden Analysen. Der Nachteil ist, dass der Messfehler deutlich verstärkt wird.

Die Paretoskalierung ist der Unitvariance-Methode sehr ähnlich, allerdings wird hier auf die Wurzel der Standardabweichung skaliert.

$$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_i}{\sqrt{\sigma_i}}$$

Dadurch werden Variablen mit großen Werten aber kleiner Streuung stärker abgeschwächt als Variablen mit kleinen Werten aber im Verhältnis großer Streuung. Außerdem bleibt die Dimension der Daten erhalten. Die Daten bleiben näher an der ursprünglichen Werteverteilung, allerdings ist die Methode auch empfindlicher gegenüber großen Veränderungen.

Die Methoden der statistischen Datenauswertung lassen sich in zwei prinzipielle Kategorien unterteilen: univariat und multivariat. Ersteres bedeutet, jede Variable wird für sich, letzteres mehrere oder alle Variablen werden im Zusammenhang betrachtet. Univariat lässt sich jede Variable durch einige Kenngrößen beschreiben. Die charakteristischen Größen sind zum einen Mittelwert, Median und Modalwert und zum anderen Varianz und Standardabweichung. Diese Daten können in Boxplots sehr übersichtlich dargestellt werden, meist sind zusätzlich Quantile sowie Range und Interquartilbereich angegeben. Jeder dieser Werte besitzt eine gewisse Aussagekraft, ist aber auch stark beeinflussbar und muss deshalb vorsichtig betrachtet werden. Der Mittelwert μ wird durch Extremwerte, egal ob „reale" Werte oder Ausreißer, beeinflusst, da jeder Wert gleich stark gewichtet wird. Der Median ist hier robuster, da er den mittleren aller Werte auf deren Gesamtanzahl bezogen angibt. Der Modalwert schließlich ist der Wert, der am häufigsten vorkommt. Sind die Daten perfekt normalverteilt, so liegen alle drei Werte aufeinander. Je unregelmäßiger und asymmetrischer jedoch die Verteilung, umso weiter entfernen sie sich voneinander. Die Varianz gibt die Abweichung vom Mittelwert an, sie berechnet sich als Quadrat des durchschnittlichen Abstands der Werte vom Mittelwert μ. Die Standardabweichung wiederum ist die Quadratwurzel der Varianz. 1σ enthält 68% der Daten, also der Bereich von -1 σ bis +1 σ, 2 σ 95% der Daten, 3 σ 99% der Daten. Das erste Quantil entspricht dem Wert, bei dem 25% der Daten kleiner sind, das zweite Quantil entspricht dem Median, also sind 50% der Daten größer und 50% kleiner und 25% der Daten sind größer als das dritte Quantil.

Zwischen dem ersten und dritten Quantil liegt also die Hälfte der Daten. Der Range ist die Spannweite der Daten, das heißt Minimum bis Maximum.

Multivariate Verfahren betrachten nicht jeweils nur eine Variable, sondern die gesamte Datenmatrix. Dadurch kann die Struktur der Daten entschlüsselt werden, wohingegen mit univariaten Methoden zwar zu jeder Variable Informationen erhalten werden, jedoch das Gesamtbild und damit die zugrundeliegenden Muster verborgen bleiben. Multivariat bedeutet in diesem Falle, dass alle Variablen gleichzeitig betrachtet und analysiert werden. Dadurch können Gruppen innerhalb der betrachteten Population erkannt werden. In der multivariaten Datenanalyse wird nochmals in überwachte und unüberwachte Methoden unterschieden. Für letztere ist keinerlei Vorkenntnis über die Proben und damit keine klar definierte Fragestellung nötig. Das heißt die Ergebnisse können nahezu völlig frei interpretiert werden. Dies erfordert allerdings die Möglichkeit, statistische Daten interpretierbar darzustellen. Multivariate Datenanalyse gibt zum einen Aufschluss über Verbindungen zwischen den Variablen, das heißt Muster, die die Spalten (also Variablen) der Matrix miteinander verbinden. Zum anderen gibt sie Information über Gruppen und deren „Entfernung" voneinander. Verschiedene Größen beschreiben dies.

Das große Problem multivariater Datenmatrizen ist die große Anzahl Variablen und damit Dimensionen, die gleichzeitig betrachtet werden. Eine Vereinfachung der Daten ist notwendig, um diese auswerten zu können. Mit der Hauptkomponentenanalyse (Principal Component Analysis, PCA), die eine Linearkombination der Variablen darstellt, ist das möglich. Die Hauptkomponenten (PCs) sind unkorreliert und nach ihrem Beitrag, um die Variation zu erklären, geordnet. Im Idealfall wird mittels PCA eine kleine Anzahl von Variablen erzeugt, mit denen der Datensatz visualisiert werden kann. Nicht immer ist die erste Komponente die für die Studie interessanteste, da sie oft durch besonders intensive Metabolite bestimmt wird, oft wird erst mit der zweiten oder auch dritten Hauptkomponente Aufschluss über niedriger konzentrierte Substanzen erhalten. Allerdings muss bedacht werden, dass PCA im Falle unkorrelierter Variablen keinerlei brauchbare Ergebnisse liefert. Die erste Hauptkomponente y_1, also die mit der größten Varianz, ist definiert als:

$$y_1 = a_{11}x_1 + a_{12}x_2 + \ldots + a_{1q}x_q$$

mit den korrelierten Variablen xi und den Koeffizienten a_{ji}. Die zweite Hauptkomponente, mit der zweitgrößten Varianz, ist folglich:

$$y_2 = a_{21}x_1 + a_{22}x_2 + \ldots + a_{2q}x_q$$

Die Gesamtvarianz der q Hauptkomponenten entspricht der Gesamtvarianz der ursprünglichen Datenmatrix. Die Koeffizienten a_{ji} sind so gewählt, dass die neu berechneten Variablen, im Gegensatz zu den Ursprungsdaten, nicht korreliert sind. Geometrisch betrachtet ist die erste Hauptkomponente die Gerade mit dem besten Fit an die q-dimensionalen Beobachtungen in der Probe. Im Scoresplot wird jede Probe als Punkt dargestellt, ihre Position ergibt sich aus den ersten beiden Hauptkomponenten. Prinzipiell kann natürlich jede beliebige Kombination von Hauptkomponenten als Graph dargestellt werden, oft werden auch dreidimensionale Diagramme verwendet. Diese Darstellung ermöglicht Gruppierungen zu erkennen und zu überprüfen, ob sie mit den tatsächlichen Untergruppen der Messreihe übereinstimmen, sofern diese bekannt sind. Die zweite häufig verwendete Darstellung ist der Loadingsplot, er zeigt den Beitrag jeder Variablen zu der jeweiligen Hauptkomponente. Daraus ergibt sich die zweite wichtige Information: welche Metabolite sind maßgeblich für die Trennung verantwortlich?

Neben unüberwachten Methoden wie PCA werden überwachte Methoden wie die „Partial Least Square Discriminant Analysis" (PLS-DA) verwendet. PLS-Regression wird verwendet, um die Verbindung einer großen Anzahl Variablen und der Gruppenzugehörigkeit der Probe (Responsevariable) zu modellieren. Die Daten werden in ein Trainingsset und ein Testset unterteilt, mit ersterem wird das Modell erstellt, das mit letzterem validiert wird. Es wird ein Set neuer Variablen erstellt, das aus den X-Scores (Predictors) der Y-Variablen (Responses, Eigenschaften) besteht. Wie bei der PCA sind die neuen Variablen Linearkombinationen der ursprünglichen. Der Unterschied zur PCA ist, dass in ersterer die Variablen so gewählt werden, dass die größte Variation zwischen den Proben beschrieben wird, völlig unabhängig von einem möglichen Zusammenhang zwischen den Proben. In der PLS werden die Variablen, die eine hohe Korrelation mit der Responsevariable zeigen stärker gewichtet, da diese die Vorhersage am ehesten ermöglichen. Die Darstellung der Daten erfolgt analog zur PCA, das heißt, im Scoresplot werden die relativen Positionen der Proben zueinander betrachtet, im Loadingsplot welche Peaks in welchem Maße dazu beitragen.

Fragestellung

Können verschiedene Kaffeeproben anhand der Spektren der wässrigen Extrakte voneinander unterschieden werden? Welche Eigenschaft (z.B. Röstung, Bohne, Herkunft etc) führt zu den stärksten Unterschieden? Welche Metabolite sind maßgeblich für die Unterscheidung?

Literatur

Metabolomics:

Metabolic profiling, metabolomic and metabonomic procedures for NMR spectroscopy of urine, plasma, serum and tissue extracts

Olaf Beckonert, Hector C Keun, Timothy M D Ebbels, Jacob Bundy, Elaine Holmes, John C Lindon & Jeremy K Nicholson, Nature Protocols 2692 – 2703 (2007); doi:10.1038/nprot.2007.376

Systems biology: Metabonomics

Jeremy K. Nicholson & John C. Lindon, Nature 455, 1054-1056 (23 October 2008); doi:10.1038/4551054a

Innovation: Metabolomics: the apogee of the omics trilogy

Gary J. Patti[1], Oscar Yanes[2] & Gary Siuzdak[3], Nature Reviews Molecular Cell Biology 13, 263-269 (April 2012), doi:10.1038/nrm3314

NMR:

H. Friebolin, Ein- und zweidimensionale NMR-Spektroskopie, 3. Aufl., Wiley-VCH, Weinheim, 1999

Aufgaben (die Beantwortung der Aufgaben sind Bestandteil des Kolloquiums)

- Welche anderen Methoden können für Metabolomics-Studien eingesetzt werden? Was sind hier Vor- und Nachteile der NMR-Spektroskopie?
- Warum spielt die Reproduzierbarkeit der Daten eine so große Rolle?
- Warum muss ein Puffer verwendet werden?
- Wo sind Unterschiede zwischen den verschiedenen Kaffeeproben zu erwarten?
- Was bedeutet univariat, was multivariat?
- Welche Methoden werden wir in der statistischen Datenauswertung verwenden? Was sagen diese aus?
- Was ist der Unterschied zwischen überwachten und unüberwachten Methoden der statistischen Datenauswertung?
- Was ist eine Hauptkomponentenanalyse?

Versuchsdurchführung

Tag 1: Probenvorbereitung und Messung

Material

Jedes Gruppenmitglied bringt – wenn möglich – zwei Proben gemahlenen Kaffee mit. Hersteller, Herkunft der Bohnen, Mischverhältnis von Arabica und Robusta sowie Röstgrad sollten soweit bekannt notiert werden, da diese Angaben für die Auswertung der Spektren benötigt werden.

Puffer: 200 mM NaH_2PO_4, 2 mM NaN_3, 1 mM TSP in D_2O, pH 6.0 der pH-Wert wird mit KOH und HCl eingestellt. Der Puffer wird von der ersten Gruppe für alle anderen vorbereitet. Die anderen Gruppen berechnen die benötigten Mengen der Bestandteile um 10 mL Puffer herzustellen.

Von jeder Kaffeeprobe werden Triplikate angefertigt.

Probenvorbereitung

- 0.1 g Kaffee mit 1.5 mL H_2O_{dest} mischen (vortex)
- Inkubation für 30 min bei 95 °C (Eppis nicht ganz schließen!)
- Proben bei Raumtemperatur etwas abkühlen lassen, dann 10 min bei 4 °C in die Zentrifuge stellen.
- 20 min bei höchster Stufe zentrifugieren
- 1 mL des Überstandes in ein neues Eppi überführen
- Eppis mit Parafilm verschließen, in diesen mit einer Nadel einige Löcher stechen
- Proben in flüssigem Stickstoff einfrieren
- Über Nacht lyophilisieren

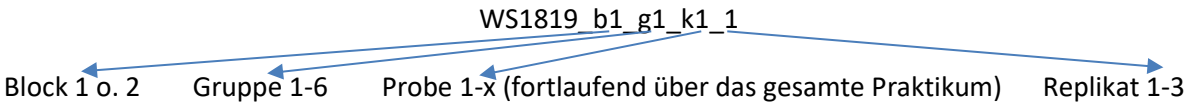
--- der nächste Schritt findet im Anschluss an einen der folgenden Versuche statt, bitte im Plan nachschauen ----

- Trockene Proben in 600 µL Puffer aufnehmen (gut vortexen!)
- 10 min bei höchster Stufe zentrifugieren
- 550 µL des Überstands in NMR-Röhrchen überführen

--- Die fertigen Proben werden vom Assistenten gemessen. Die Auswertung der Spektren wird an einem der Folgetage stattfinden (siehe Plan). ----

Tag 2: Datenauswertung

Die Spektren wurden folgendermaßen benannt:

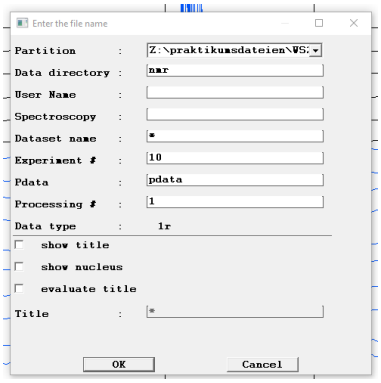


Metabolitenidentifikation und qualitativer Vergleich der Spektren

Öffnen Sie ein beliebiges Spektrum Ihrer Gruppe in Topspin.

1. Welche Metabolite sind zu erwarten?
2. In welchem Bereich im Spektrum sind aromatische Protonen zu sehen, wo Zucker, wo Aminosäuren?
3. Wo werden Signale von Koffein, Trigonellin, 16-O-Methylcafesterol und Kahweol erwartet?
4. Suchen Sie die entsprechenden Substanzen in der HMDB (hmdb.ca) und ordnen Sie die entsprechenden Signale im Spektrum zu.

Öffnen Sie alle Spektren der Gruppe in Amix:
Amix-Viewer -> File -> Open TOPSPIN 1D file
Partition: Z:\praktikumsdateien\WS201819
alle anderen Felder siehe Abbildung
Dann die Spektren der Gruppe auswählen.



Färben Sie die einander zugehörigen Replikate entsprechend ein:
Rechtsklick auf die Spektren -> color by group
Zahlen entsprechend eintragen.
unter Config -> Display -> display file names of spectra:
short filenames ausäwhlen, falls die Namen der Spektren nicht bereits angezeigt werden.

1. Wie unterscheiden sich die Proben, welche Gemeinsamkeiten gibt es?
2. Unterscheiden sich die Spektren der gleichen Probe? Worauf sind mögliche Unterschiede zurückzuführen?

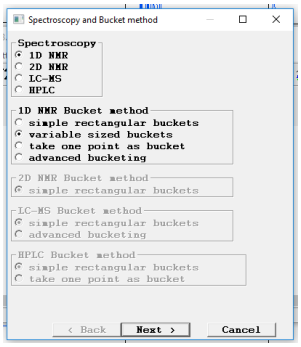
Vorbereitung der statistischen Datenauswertung

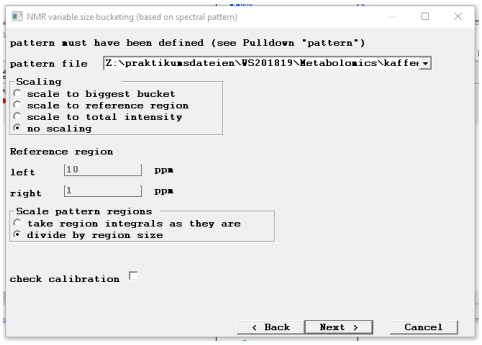
Öffnen Sie das Patternfile „praktikum_ws1819“ und überprüfen Sie, ob die gewählten Abschnitte zu den Signalen der Spektren passen:

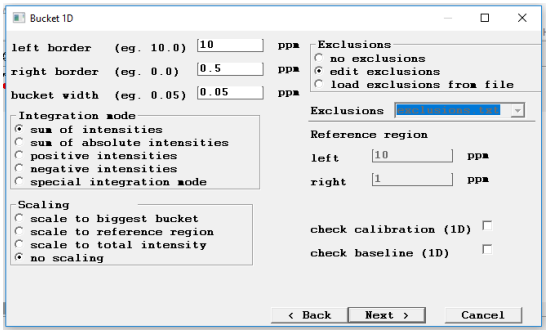
Patterns -> Open pattern file -> Z:\praktikumsdateien\WS201819\Metabolomics\kaffee

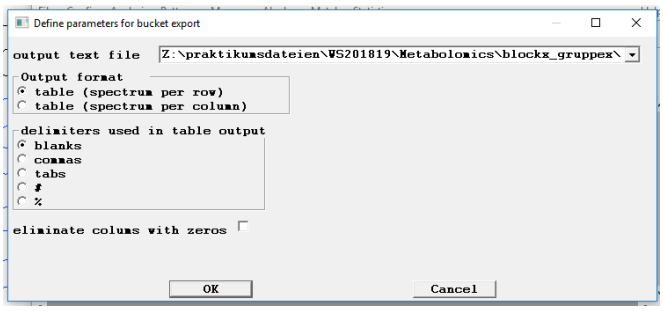
Erstellen Sie mit diesem Patternfile aus allen Spektren des Praktikumsblocks eine bucket table:

1. Amix-Tools -> Buckets, Statistics -> neues Fenster öffnet sich:
2. Statistics -> Bucket table -> New → „1D NMR“ und „variable sized buckets“ auswählen
3. Next → pattern file eintragen, “no scaling”, “divide by region size”
4. Next → bucket table folder:
   Z:\praktikumsdateien\WS201819\Metabolomics\blockx_gruppex\...
5. Next → Data sources: “Topspin data tree”
6. OK → Spektren auswählen wie vorher
7. Select next “nein”
8. Statistics -> bucket table -> export: → output text file in Ihren Ordner speichern
9. Wiederholen Sie Schritt 2, wählen Sie aber diesmal “simple rectangular buckets” aus.
10. Next → Bereich von 10 bis 0.5 ppm auswählen, bucket width 0.05 ppm, edit exclusions auswählen
11. Next → Bereich des Wassersignals ausschließen
12. Next: → weiter mit Schritt 4 bis 8 (Ordner- bzw Dateinamen entsprechend anpassen)









Um die bucket tables in Metaboanalyst nutzen zu können müssen sie in Excel angepasst werden. Öffnen Sie dafür die entsprechende Datei mit Excel und bringen Sie sie in das in der Abbildung vorgegebene Format.

	A	B	C	D	E
1	sample name	class	v1	v2	v3
2	spectrum_1	c1	xxx	xxx	xxx
3	spektrum_2	c2	xxx	xxx	xxx
4	Spektrum_3	c1	xxx	xxx	xxx
5	Spektrum_4	c2	xxx	xxx	xxx
6					

„class“ entspricht der Gruppeneinteilung, die man später in der Statisitik haben möchte. Also Beispielsweise Robusta/Arabica, Espresso/Filterkaffee, Praktikumsgruppe etc.
Sample name ist der individuelle Name jedes Spektrums. Da nur eine begrenzte Zahl Zeichen angezeigt werden kann sollte er so kurz wie möglich gewählt werden.
Speichern Sie die Tabelle schließlich im „.csv“-Format ab.

Die Informationen zu den Proben sind in der Datei „samples“ im Praktikumsordner zu finden.

Statistische Datenauswertung mit Metaboanalyst.ca

Die Auswertung wird für beide bucket tables durchgeführt. Je nach Vorbereitung der Datenmatrix kann es notwendig sein, die Einteilung der „classes“ nachträglich zu ändern und die Analyse nochmals durchzugehen. Außerdem sollen die verschiedenen Normalisierungs- und Skalierungsmethoden verglichen werden. Ziel der statistischen Auswertung ist es, zu vergleichen, mit welchen Parametern aufgrund der vorhandenen Daten eine Aussage getroffen werden kann. Des Weiteren soll kritisch betrachtet werden, welche Methoden wie und vor allem wie stark die Daten manipulieren und wie viel Aussagekraft dem jeweiligen Ergebnis dann noch zugesprochen werden kann.

Vorbereitung der Daten für die eigentliche Statistik

1) MetaboAnalyst (www.metaboanalyst.ca) wird aufgerufen.
2) >> click here to start <<
3) Statistical Analysis
4) Upload your data:
   Data Type: Spectral bins
   Format: Samples in rows (unpaired)
   Data File: Das entsprechende .csv-file
   -> Submit
5) Data Integrity Check:
   Missing value estimation ist nicht nötig
   -> Skip
6) Data filtering:
   None
   -> Proceed
7) Sample normalization:
   Hier können verschiedene Varianten ausprobiert werden:
   a) None
   b) by sum – Teilt durch die Summe aller buckets eines Spektrums
   c) by median – Subtrahiert den Median aller buckets eines Spektrums vom jeweiligen bucketwert
8) Data transformation:
   None
9) Data scaling:
   Hier können verschiedene Varianten ausprobiert werden:
   a) None
   b) Mean centering
   c) Auto scaling (auch unitvariance scaling genannt)
   d) Pareto scaling
      -> Normalize, dann proceed

Jetzt folgt die eigentliche Statistik

1) ANOVA (Analysis of variance)
   a) Welche Variablen sind signifikant, welche nicht? Ist dies auch in den jeweiligen Boxplots ersichtlich?
2) Correlation Analysis
   a) Welche Variablen sind positive, welche negative korreliert. Was könnte das bedeuten?
3) PCA
   a) Scree Plot: Kann mit der PCA die Varianz der Proben ausreichend gut erklärt werden?
   b) 2D Scores Plot: Ist eine Trennung der verschiedenen Proben sichtbar? Entspricht diese den gewählten Gruppen oder sind möglicherweise andere Eigenschaften maßgeblicher?
   c) 3D Scores Plot: Wird die Trennung deutlicher, wenn die dritte PC dazu genommen wird oder ändert sich nichts?
   d) Loadings Plot: Welche Variablen sind vor Allem für die Trennung verantwortlich (am weitesten vom Nullpunkt entfernt)? Sind alle einer Substanz zugeordneten buckets gleich wichtig für die Trennung? Wie sehen die zugehörigen Boxplots aus?
4) PLS-DA
   a) 2D Scores Plot: Kann die PLS-DA die Gruppen gut voneinander trennen? Sieht die Trennung anders aus als die der PCA? Ist die Trennung im Vergleich zur PCA deutlicher?
   b) 3D Scores Plot: Wird die Trennung deutlicher, wenn die dritte Komponente dazu genommen wird oder ändert sich nichts?
   c) Loadings Plot: Welche Variablen sind vor Allem für die Trennung verantwortlich (am weitesten vom Nullpunkt entfernt)? Sind alle einer Substanz zugeordneten buckets gleich wichtig für die Trennung? Wie sehen die zugehörigen Boxplots aus?
      Sind die gleichen Variablen besonders wichtig wie in der PCA?
5) Dendrogram
   a) Sind die Replikate der gleichen Probe nah beieinander oder gibt es Ausreiser?
   b) Wie werden die Proben relativ zueinander eingeteilt, welche sind sich nah, welche sind besonders weit voneinander entfernt?
6) Download -> Generate report -> Analysis report
   erzeugt ein PDF, in dem die wesentlichen Ergebnisse der Analyse zusammengefasst sind.
   Vorsicht: Die Grafiken sind nicht immer so verwendbar, wie sie im Report ausgegeben werden. Beispielsweise sind die Sample names in PCA und PLS-DA nicht immer angezeigt und die Loadings sind ebenfalls nicht immer beschriftet. Histogramme sind generell nicht enthalten. Deshalb bitte direkt überprüfen, ob die für's Protokoll benötigten Abbildungen so vorhanden sind, wenn nicht direkt im Metaboanalyst abspeichern.

Teil 2:

Planen Sie einen Metabolomicsversuch. Die Themen werden am Versuchstag beknnat gegeben.

Geben Sie an, wie die Proben vorbereitet werden (Extraktionsmethode, Lösungsmittel, Puffer, pH) und auf was unter Umständen besonders zu achten ist. Begründen Sie jede Angabe. Welche Probleme könnten bei der Probenvorbereitung auftreten? Was könnte bei der Messung stören? Werden die für die Fragestellung interessanten Metabolite mit der gewählten Extraktionsmethode erfasst? Welche Metabolite sind vermutlich zu erwarten? Wie würden Sie bei der statistischen Datenauswertung vorgehen (Normierung/Skalierung)?