Metabolomics

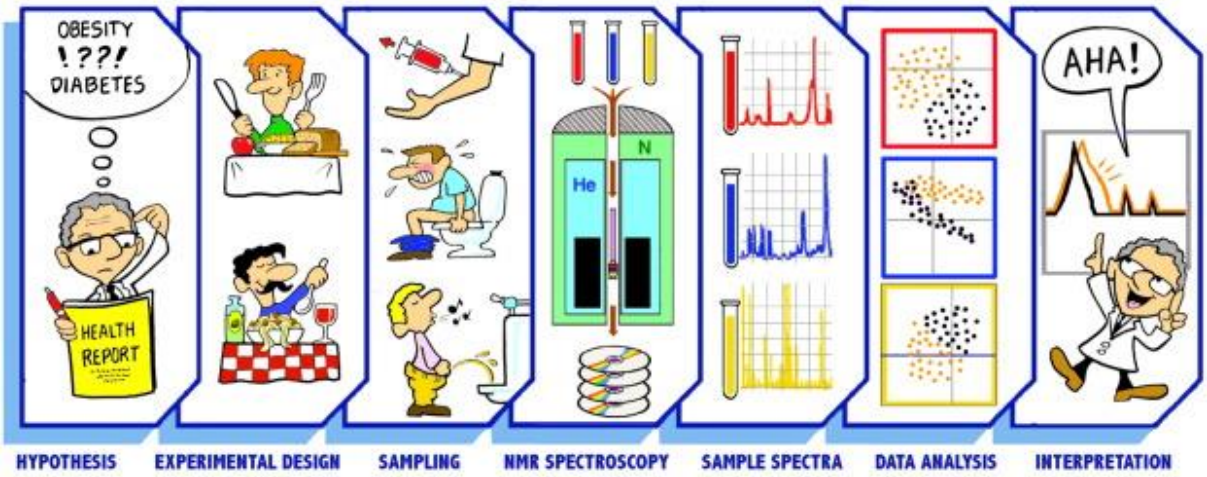
"Quantitative measurement of the dynamic multiparametric metabolic response of living systems to pathophysiological stimuli or genetic modification." (Jeremy Nicholson)

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, Identifizierung und Quantifizierung von Metaboliten in Organismen, Geweben, Sekreten oder Biofluiden etc..

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. Des Weiteren kommt die Metabolomicsforschung in der Lebensmittelwissenschaft zum Einsatz, beispielsweise im Bereich der Qualitätskontrolle

In diesem Praktikum werden wir eine kurze Einführung in die wichtigsten Aspekte bezüglich der Extraktion und Probenvorbereitung am Beispiel 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 proportional 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. Damit impliziert man also, dass in jeder Probe die gleiche Gesamtmenge dieser Substanzen vorhanden sein muss. Es gibt noch andere Normierungsmethoden, wie zum Beispiel die Normierung auf den Median. Bei dieser Methode wird der Median aller Buckets eines Spektrums 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) in 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. Durch die Skalierung werden kleine Werte allerdings 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, dass jede Variable für sich betrachtet wird, letzteres, dass mehrere oder alle Variablen im Zusammenhang betrachtet werden. 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 unüberwachte (‘unsupervised‘) und überwachte (‘supervised‘) Methoden unterschieden. Für erstere 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 x_i 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 Gruppen 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. Da die Klassenzugehörigkeit der Proben in die Auswertung mit einbezogen wird, kommt es typischerweise zu einer deutlicheren Trennung der Gruppen. Dadurch können die Daten detaillierter betrachtet werden, allerdings kann auch eine rein zufällige Trennung der Gruppen entstehen. 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.

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

Statistische Datenauswertung:

- **MetaboAnalyst: a web server for metabolomic data analysis and interpretation**
  Jianguo Xia, Nick Psychogios, Nelson Young, David S. Wishart, Nucleic Acids Research, Volume 37, Issue suppl_2, 1 July 2009, Pages W652–W660, https://doi.org/10.1093/nar/gkp356
- **Centering, scaling, and transformations: improving the biological information content of metabolomics data**
  van den Berg, R.A., Hoefsloot, H.C., Westerhuis, J.A. et al., BMC Genomics 7, 142 (2006). https://doi.org/10.1186/1471-2164-7-142

Probenvorbereitung

Die Kaffeeproben werden <u>während des Massenspektrometrie-Versuchs</u> vorbereitet. Dabei stellt jeder Student drei Proben eines Kaffees her.

<u>Extraktionsprotokoll:</u>

1. Einwaage von 0.1 g des jeweiligen Kaffees (Genaue Einwaage notieren)
2. Mischen mit 1.2 mL H_2O_{bidest} (vortexen)
3. Inkubation für 10 min bei 95 °C
4. Kurz abkühlen lassen, dann 15 min im Kühlschrank runterkühlen
5. Zentrifugation für 15 min bei maximaler Geschwindigkeit und 4°C
6. 630 µL des Überstandes mit 70 µl 10x Puffer mischen (vortexen)
7. 600 µL in NMR-Röhrchen transferieren

10x Puffer: 2 M NaH_2PO_4, 2 mM NaN_3, 1 mM TSP in D_2O, pH 6

Die Proben werden anschließend von der Assistentin gemessen und die Daten aller Gruppen den Studenten für den Metabolomics-Versuch zur Verfügung gestellt.

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 im Vergleich zur Massenspektrometrie?
- Warum spielt die Reproduzierbarkeit der Daten eine so große Rolle? Wie kann eine möglichst hohe Reproduzierbarkeit erreicht werden.
- Warum muss ein Puffer verwendet werden? Welche Puffersubstanz bietet sich für die NMR-Spektroskopie an? Wofür dienen die einzelnen Bestandteile des verwendeten Puffers?
- Welchen Zweck hatten die einzelnen Schritte der Probenvorbereitung?
- Bei der Vorbereitung von Metabolomics-Proben wird häufig ein Lyophilisierungs-Schritt durchgeführt. Was ist das Prinzip dieses Verfahrens und warum wird es verwendet?
- Eine der Kaffeeproben ist entkoeffiniert. Welche Verfahren gibt es?
- Wo sind Unterschiede zwischen den verschiedenen Kaffeeproben zu erwarten?
- Was bedeutet Normierung, was Skalierung. Welche Methoden gibt es?
- Was bedeutet univariat, was multivariat?
- Was ist ein Boxplot, was sagt er aus?
- 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?

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?

- Behaltet bei der Versuchsdurchführung immer die übergreifende Fragestellung im Hinterkopf
- Orientiert euch beim Schreiben des Protokolls an den Fragen der Versuchsdurchführung. Beantwortet bitte auch die zusätzlich angegebenen Fragen.
- Speichern alle Abbildungen, die ihr später für das Protokoll braucht!

Datenauswertung

Die Spektren wurden in Topspin folgendermaßen benannt:

Metabolitenidentifikation und Vergleich der Spektren

Öffnet ein beliebiges Spektrum in Topspin (Die Spektren befinden sich in Topspin im Ordner WS2023_Kaffee)

1. Welche Metaboliten sind in Kaffee zu erwarten?
2. In welchem Bereich im Spektrum sind aromatische Protonen zu sehen, wo Zucker, wo Aminosäuren?
3. Wo werden Signale von Koffein und Trigonellin erwartet?
4. Sucht die entsprechenden Substanzen in der „Human Metabolome Database“ HMDB (hmdb.ca) oder der FoodDB (fooddb.ca) und ordnet die entsprechenden Signale im Spektrum zu.
5. Versucht mit Chenomx sowie mit Hilfe der Datenbanken weitere Metaboliten zu identifizieren.

- *Im Protokoll diskutieren: Wie kann nachgewiesen werden, dass die Zuordnung der Substanzen zu den Peaks richtig ist?*

Öffnet die weiteren Spektren des gleichen Kaffees (Topspin Kommando „.md“ → Multidisplay Ansicht öffnet sich und es können mehrere Spektren in ein Fenster gezogen werden)

6. Unterscheiden sich die Spektren oder sind sie reproduzierbar? Worauf sind mögliche Unterschiede zurückzuführen?

Öffnet jeweils ein Spektrum eines Arabica und eines Robusta Kaffees, sowie des koffeinfreien Kaffees. Die Informationen zu den Kaffees sind in der Datei „WS2023_kaffee_samples.xlsx“ enthalten.

7. Wie unterscheiden sich die Proben? Welche Gemeinsamkeiten fallen auf? Gibt es Metaboliten, die besonders hervorstechen? Lässt sich anhand der Spektren der koffeinfreie Kaffee identifizieren?
8. Berechnet den Koffeingehalt der drei Kaffees:
    - Tipp: Erinnert euch an den ersten NMR-Versuch (Berechnung Methanol- und Ethanolanteil). Im Gegensatz dazu kennen wir jetzt aber die Konzentration einer Substanz (welche?), die wir als absolute Referenz verwenden können.
    - Es reicht, wenn ihr jeweils ein Koffeinsignal integriert. Sucht euch am besten eins aus das möglichst isoliert von anderen Substanzen ist
    - ➢ *Im Protokoll diskutieren: Stimmen die berechneten Werte mit Literaturangaben überein?*

Vorbereitung der statistischen Datenauswertung

Die Buckettable wird vom Assistenten zur Verfügung gestellt („ws2023_kaffee_buckettable.xlsx“). Öffnet die Tabelle zuerst in Excel und macht euch mit der Datenstruktur vertraut.

	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

„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. „class“ entspricht der Gruppeneinteilung, die man später in der Statistik haben möchte. Für die ersten Schritte sind die Spektren nach Kaffeeprobe eingeteilt. Später können die Proben hier beispielsweise in Robusta/Arabica, Espresso/Filterkaffee etc. aufgeteilt werden. Die restlichen Spalten enthalten die Integrale für die entsprechenden Buckets.

9. Einige Buckets wurden bereits mit den entsprechenden Namen der Substanzen benannt. Konntet ihr noch mehr Substanzen identifizieren? Falls ja können die zugehörigen Buckets zusammen mit dem Assistenten benannt werden.

Speichert die Tabelle anschließend im „.csv“-Format ab.

Statistische Datenauswertung mit Metaboanalyst.ca

Ziel der statistischen Auswertung ist es, zu vergleichen, mit welchen Parametern anhand der vorhandenen Daten eine Aussage getroffen werden kann. Des Weiteren soll kritisch betrachtet werden, welche Normierungs- und Skalierungsmethoden wie und vor allem wie stark die Daten manipulieren und wie viel Aussagekraft dem jeweiligen Ergebnis dann noch zugesprochen werden kann.

Alle Graphiken aus MetaboAnalyst können gespeichert werden durch Rechtsklick -> Grafik speichern unter, oder klicken auf dieses Symbol: 

Hochladen der Daten in MetaboAnalyst (Programm für Metabolomics Datenanalyse)

- MetaboAnalyst (www.metaboanalyst.ca) wird aufgerufen.
- >> click here to start <<
- Statistical Analysis [one factor]
- Upload your data (A plain text file (.txt or .csv)):
  Data Type: Spectral bins
  Format: Samples in rows (unpaired)
  Data File: Das entsprechende .csv-file
  -> Submit
- Data Integrity Check:
  Missing value estimation ist nicht nötig
  -> Proceed
- Data filtering:
  None
  -> Proceed

Normalisierung und Skalierung der Daten:

Nun sollen verschiedene Normalisierungs- und Skalierungsmethoden verglichen werden:

- Sample normalization:
  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
- Data transformation:
  None
- Data scaling:
  a) None
  b) Auto scaling (auch unitvariance scaling genannt) – Zentrierte Daten werden durch die Standardabweichung geteilt
  c) Pareto scaling: – Zentrierte Daten werden durch die Wurzel der Standardabweichung geteilt

-> Normalize

Es gibt verschiedene Möglichkeiten, die Methoden zu vergleichen:

- Über die Dichteverteilung, hier kann überprüft werden wie normalverteilt die Daten sind (-> View Result)

- Oder über den 2D Scores- sowie Loadingsplot der PCA (-> Proceed -> Principal Component Analysis (PCA))

  Kurze Anmerkung zur Darstellung der PCA: Beim Scores Plot können über „Display Sample Names -> Update“ die Probennamen direkt im Plot angezeigt werden. Beim Loadings Plot kann das selbe über „Label all variables -> Update“ gemacht werden, allerdings werden die Namen hier erst angezeigt, wenn das Bild über  *erstellt wird.*

10. Vergleicht die nicht-normierten Daten mit den beiden oben genannten Normierungsmethoden (Eine Skalierung wird für diesen Schritt noch nicht ausgeführt). Muss bei der PCA der Scores- oder Loadingsplot betrachtet werden? Wie verändert sich dieser?
    → Welche Variante erscheint am sinnvollsten und warum?
11. Führt die nächsten Schritte mit der gewählten Normierungsmethode durch
    Vergleicht die unskalierten Daten mit den beiden oben genannten Skalierungsmethoden. Muss bei der PCA der Scores- oder Loadingsplot betrachtet werden? Wie verändert sich dieser?
    → Welche Variante erscheint am sinnvollsten und warum?

Folgende Fragen sollen nun untersucht werden

Wählt zuerst die zuvor festgelegte Normierungs- und Skalierungsmethode aus.

12. PCA (2D Scores Plot)
    - Ist eine Gruppierung der verschiedenen Proben erkennbar? Wonach gruppieren sie sich?
13. Dendrogramm
    - Sind die Replikate der gleichen Probe nah beieinander oder gibt es Ausreiser?
    - Wie werden die Proben relativ zueinander eingeteilt, welche sind sich nah, welche sind besonders weit voneinander entfernt? Lassen sich die „Arme“ einteilen in beispielsweise Arabica/Robusta und Filter/Espresso?
14. Da Arabica-Bohnen teurer sind als Robusta-Bohnen, werden Arabica-Kaffees häufig mit Robusta aufgefüllt ohne dass dies angegeben wird. Deshalb wird zur Qualitätskontrolle in Laboren anhand verschiedener Marker untersucht, ob es sich dabei wirklich um 100% Arabica Kaffee handelt.
    Eure Probe „U1“ wurde vom Hersteller auch als 100% Arabica deklariert. Würdet ihr anhand der PCA und des Dendrogramms sagen, dass das auch stimmt? Allerdings war nicht angegeben, ob es sich um einen Espresso oder Filterkaffee handelt. Was meint ihr?

Als nächstes soll die Einteilung der „class“ geändert werden. Teilt die Proben dafür entsprechend einer Klassifizierung ein (z.B. Arabica vs. Robusta **oder** Filter vs. Espresso). Anhand des zuvor betrachteten Dendrogramms könnt ihr abschätzen, zwischen welchen Klassen der Unterschied am deutlichsten ist. Hier können einzelne Proben, die die Analyse verzerren könnten entfernt werden (beispielsweise der entkoffeinierte Kaffee, oder auch die unbekannte Probe).

Mithilfe verschiedener statistischer Methoden soll nun analysiert werden, anhand welcher Substanzen sich die gewählten Gruppen voneinander unterscheiden lassen

Ladet die Daten wie zuvor in MetaboAnalyst hoch und wählt die zuvor bestimmte Normierungs- und Skalierungsmethode aus.

Univariate Analyse:

15. Entweder T-Test (für 2 Gruppen) oder ANOVA (Analysis of variance, für mehr als 2 Gruppen)
    - Welche Variablen sind signifikant, welche nicht? Ist dies auch in den jeweiligen Boxplots ersichtlich?

Multivariate Analyse:

16. PCA
    - Scree Plot: Kann mit der PCA die Varianz der Proben ausreichend gut erklärt werden?
    - 2D Scores Plot: Ist eine Trennung der verschiedenen Proben sichtbar? Entspricht diese den gewählten Gruppen oder sind möglicherweise andere Eigenschaften maßgeblicher? Ist zusätzlich zu den gewählten Gruppen noch eine weitere Aufteilung der Proben sichtbar?
    - 3D Scores Plot: Wird die Trennung deutlicher, wenn die dritte PC dazu genommen wird oder ändert sich nichts?
    - 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? Falls das nicht der Fall ist – was könnte der Grund dafür sein? (Im Biplot könnt ihr außerdem den Scores Plot und Loadingsplot zusammen in einer Abbildung sehen)
17. PLS-DA
    - 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?
    - 3D Scores Plot: Wird die Trennung deutlicher, wenn die dritte Komponente dazu genommen wird oder ändert sich nichts?
    - Loadings Plot: Welche Variablen sind vor allem für die Trennung verantwortlich (am weitesten vom Nullpunkt entfernt)? Sind die gleichen Variablen besonders wichtig wie in der PCA?

- *Anhand welcher Substanzen lassen sich die gewählten Gruppen unterscheiden? In welcher Gruppe liegen die Substanzen in höherer bzw. niedrigerer Konzentration vor? Diskutiert im Protokoll auch, ob die gefundenen Unterschiede Literaturangaben entsprechen*