Mikrobiomanalyse unter Berücksichtigung biologischer Datenstruktur

Dissertation

zur Erlangung des akademischen Grades

doctor rerum naturalium
(Dr. rer. nat)

von Dipl.-Biol. Dipl.-Math. Kai Lars Antweiler
geb. am 21.06.1977 in Köln

genehmigt durch die Fakultät für Mathematik
der Otto-von-Guericke-Universität Magdeburg

Gutachter:

Prof. Rainer Schwabe
apl. Prof. Ekkehard Glimm
apl. Prof. Siegfried Kropf

eingereicht am: 29.06.2021
Verteidigung am: 16.11.2021

Inhaltsverzeichnis

Zusammenfassung

Das Ziel dieser Arbeit ist biologische Konzepte in die Datenauswertung von Mikrobiomstudien zu integrieren.

Mikrobiomanalysen vergleichen relative Häufigkeiten der Organismen aus Mikrobengemeinschaften verschiedener Proben – z.B. Proben aus Wurzelbereichen unterschiedlich behandelter Pflanzen. Diese Häufigkeitsdaten sind hochdimensional und dünnbesetzt. Der Stichprobenumfang an Pflanzen ist meist gering. Aus den Daten werden oft abstrakte Abstände zwischen Pflanzen berechnet und mit Permutationstests auf Unterschiede geprüft.

Abstände lassen sich auch für Äquivalenztests nutzen, die sonst durch die Hochdimensionalität zu geringe Güte hätten. Anstatt den Anspruch zu erheben für jede Variable Äquivalenz nachzuweisen, wird diese nur für das Abstandsmaß gezeigt. Durch ökologisch-sinnhafte Abstände lassen sich geeignete Testverfahren konstruieren. Das größte Problem stellt die zuverlässige Varianzschätzung einer ggf. komplizierten Abstandsvariable bei kleinen Stichprobengrößen dar. Zur Varianzschätzung wurde in dieser Arbeit ein 2-Stichproben-Jackknife-Verfahren identifiziert und in seiner Anwendbarkeit auf unbalancierte Designs erweitert.

Das Verwenden von Verwandtschaftsbeziehungen (also Phylogenetik) zwischen Mikroben in Häufigkeitsauswertungen ist ein junges Forschungsgebiet. Zusammenhänge zwischen Genomsequenz, Phylogenetik und Messverfahren, wurden hier zusammengetragen. In dieser Arbeit wurden zwei Verfahren an phylogenetische Zusatzinformation angepasst. 1. Hierarchische Tests bilden baumförmig organisierte Variablen-Cluster, die geschickt getestet werden. Hier wird ein Baum, der die Verwandtschaften widerspiegelt, vorgegeben. Somit führen alle Ergebnisse der Prozedur zu biologisch sinnvollen Gruppen von Mikroben. Zu diesen Ergebnissen wurde hier auch eine übersichtliche Darstellungsmethode entwickelt. 2. Sequentielles Testen mit datenabhängig geordneten Hypothesen nach Kropf zeigt hohe Güte – verliert diese jedoch mit zunehmender Varianzungleichheit zwischen Variablen. Hier wird Phylogenetik verwendet, um Variablengruppen zu bilden, deren Varianzen ähnlicher sein sollten und somit die Testbedingungen verbessert.

In dieser Arbeit konnte gezeigt werden, dass Äquivalenztests in Mikrobiomanalysen sinnvoll möglich sind und auch multiple Testverfahren phylogenetische Zusatzinformation ausnutzen können. Das betrifft insbesondere die Interpretier- und Darstellbarkeit der Ergebnisse dieser Analysen. Grenzen und Möglichkeiten durch neue Messmethoden wurden betrachtet.

Abstract

The purpose of this dissertation is to integrate biological concepts into data analyses of microbiome studies.

Analyses of the microbiome compare relative abundances of organisms from microbial communities of biological samples – e.g. samples from the root region of plants that were treated under different conditions. These data are sparse and high-dimensional. Usually, the sample size of plants is small. Often, the data is transformed into abstract distances between the plants and tested for difference by permutation testing.

Distances can also be used for tests of equivalence that would otherwise suffer from high-dimensionality and show little power. Equivalence is only investigated in terms of the distance measure instead of proving equivalence for each variable. Testing procedures can be constructed upon ecologically meaningful distance measures. The biggest problem is the reliable estimation of variance, since the distance measure can be complicated and the sample size small. In this dissertation, a 2-sample Jackknife procedure was identified for estimating the variance and its applicability was expanded to unbalanced designs.

Utilizing phylogenetical relationships between microbes in analyses of abundances is a relatively new research topic. This dissertation includes literature research for relations between genome sequences, phylogenetics, and measuring procedures. Also, two statistical procedures were modified to incorporate phylogenetical information: 1. Hierarchical tests arrange clustered variables in a tree and test those aptly. Here, that tree is constructed to mirror the phylogenetical relationships. Thereby, all results are calculated for biologically meaningful groups of microbes. Also, to visualize those results clearly a procedure was developed, here. 2. Sequential testing based on variance ordering is a powerful measure that looses the more power the more the variances of variables deviate from one another. Here, phylogenetics is used to define clusters of variables. Because the variances inside each cluster are likely to be more similar to each other, those clusters provide better conditions for this testing procedure.

In this dissertation, it was shown that test of equivalence can be applied appropriately in microbiome analyses and that procedures for multiple testing can be improved by utilizing phylogenetical information. The latter is especially true for the visualization and interpretability of the results of analyses. Potentials and limits of methods for Next Generation Sequencing based studies were considered.

1. Einführende Bemerkungen

Mikrobiomanalysen sind ökologische Untersuchungen, bei denen die Zusammensetzung von Gemeinschaften einzelliger Lebewesen an bestimmten Orten (z.B. Bodenproben in der Landwirtschaft, Darm von Säugetieren in der Medizin) gemessen werden. Sie haben zu wichtigen Erkenntnissen in Bereichen ökologischer, medizinischer und agrarwissenschaftlicher Forschung geführt und die Möglichkeiten von Biotechnologie und Forensik erweitert (Bolyen *et al.* (2019)). Durch Next Generation Sequencing Messverfahren können inzwischen vergleichsweise sehr günstig und innerhalb kurzer Zeit die Identitäten (in Form von Genomabschnitten) von allen Bakterien, Archaeen, mikrobiellen Pilzen oder anderen Mikroorganismen einer Probe und die relativen Häufigkeiten, mit denen sie in ihr vertreten sind, geschätzt werden.

Statistische Verfahren wurden in der vorliegenden Arbeit besonders im Hinblick auf Versuche aus dem Bereich der Züchtungsforschung von Nutzpflanzen entwickelt und untersucht. Für diesen Anwendungsbereich lassen sich die Mikrobengemeinschaften der Wurzelregion einer Pflanze bestimmen und mit der von weiteren Pflanzen vergleichen. Die Zusammensetzung der Mikrobengemeinschaft wird also als multivariater Zufallsvektor einer Pflanze betrachtet. Sie hängt von der Pflanze, dem Bodentyp, dem Standort und weiteren Umweltfaktoren ab.

Die Arbeit stellt Verfahren zur Mikrobiomanalyse unter Berücksichtigung biologischer Datenstrukturen vor. Sie basiert auf der Mitarbeit in zwei geförderten Forschungsprojekten des Instituts für Biometrie und Medizinische Informatik und Weiterentwicklungen in den Folgejahren.

Im ersten Teil der Arbeit werden zunächst biologische Grundlagen dargestellt und dabei einige Herausforderungen aufgezeigt, welche die Möglichkeiten der statistischen Analysen limitieren. Weiterhin werden einige statistische Verfahren zusammengestellt, die im weiteren Verlauf benötigt werden. Im zweiten Teil werden dann Verfahren zum Äquivalenztest für Mikrobiomdaten entwickelt. Insbesondere wird dabei ein Jackknife-Verfahren weiterentwickelt, um flexibler auf Unbalanziertheiten im Studiendesign, die trotz sorgfältiger Planung entstehen können, reagieren zu können. Der dritte Teil betrachtet statistische Verfahren zum Test auf Unterschiede. Dieser Teil präsentiert, wie phylogenetische Information in die Analysen einbezogen werden kann, um so die Power der Tests oder die Interpretierbarkeit der Ergebnisse zu verbessern. Eine Diskussion im vierten Teil schließt die Arbeit ab. Weiterführende Ausführungen zu den Abstandsmaßen und zu den biologischen und technologischen Hintergründen werden in Anhängen bereitgestellt.

Anmerkung:

Notation 1. Da die Funktionalität abstandsbasierter Tests in der Regel nicht auf den Axiomen (positiv definit, symmetrisch und Dreiecks-Ungleichung) einer Metrik beruht, wird der Abstands-Begriff weitergefasst benutzt und oft auch "Dissimilarität" genannt. Das Wort "Abstand" wird ohne Anspruch auf Erfüllung der metrischen Axiome benutzt werden.

Notation 2. Wenn Literaturangaben am Ende eines Abschnitts nicht in einem Satz stehen, beziehen sie sich auf den gesamten Abschnitt. Anmerkungen, die nicht aus den Literaturquellen begründet sind, werden in solchen Abschnitten von doppelten Klammern umschlossen, z.B. ((Ein Beispiel für eine Anmerkung)). Ein Abschnitt kann durch einen Seitenumbruch, eine Einrückung, eine Abbildung oder eine Tabelle optisch unterbrochen sein. Die Quelle befindet sich dann am Ende des späteren Textstücks. Eine Quellenangabe direkt hinter der Überschrift einer Definition bezieht sich auf die gesamte Definition.

Teil I.

Grundlagen

2. Biologische Hintergründe

2.1. Genome

Genome sind oberflächlich betrachtet einfach. Sie sind aus wenigen Molekülen sequenziell zusammengesetzt und strukturell weitgehend einheitlich. Meist liegen sie als DNA-Strang vor.

Genome und Verwandtschaftsbeziehungen sind bei genauerer Betrachtung nicht so einfach, wie es auf den ersten Blick scheint. Die Information ist sequenziell gespeichert, wie ein digitales Foto, aber nur teilweise sequenziell organisiert. Wie sollten Genome am besten miteinander verglichen werden? Und wenn nicht das ganze Genom eines Organismus bestimmt werden soll, welcher Teil würde sich gut eignen und welche Probleme müssen für diesen Teil berücksichtigt werden? Verwendet werden standardmäßig Sequenzen, die für phylogenetische Analysen vorgeschlagen wurden und einige Nachteile für Mikrobiomanalysen aufweisen.

Weitere Einzelheiten zu den Faktoren, welche die Beurteilung der Gensequenzen und ihrer Ähnlichkeit betreffen, werden im Anhang B (Seite 107) dargestellt.

2.2. Next Generation Sequencing

Bei den hier betrachteten Next Generation Sequencing Verfahren wird ein komplementäres DNA-Fragment für DNA-Sequenzen von beispielsweise Bakterien angeboten, an das sich passende Fragmente von weitgehend allen Bakterien ähnlich gut anlagern können. Ein DNA-Fragment aus der Bodenprobe lagert sich an und wird letztendlich sequenziert. Aus nichtlinearen Kostengründen wird meist nur ein relativ kurzes Stück sequenziert. Der Vorgang wird je nach Höhe der Bezahlung einige tausend mal parallel ausgeführt und ergibt somit eine relative Verteilung der Häufigkeiten der enthaltenen Bakterien. Meist werden die Proben von wenigen Bakterienarten dominiert und enthalten viele Arten, von denen nur sehr geringe Häufigkeiten gemessen wurden.

Damit möglichst viele unterschiedliche Typen von Bakterien überhaupt gemessen werden können, muss ein DNA-Fragment verwendet werden, das evolutionär konserviert ist. Dazu werden meist Fragmente aus den Genombereichen verwendet, die für Untereinheiten von Ribosomen kodieren. Zur Unterscheidung zwischen Bakterien ist aber auch eine hohe Anzahl

an Mutationen wünschenswert – also weniger konservierte Abschnitte, die ebenfalls mitsequenziert werden sollten. Wenn man sich nicht für alle Bakterien interessiert, sondern nur für einen kleineren verwandten Anteil einer Bakterienpopulation, können direkt weniger konservierte Bereiche gewählt und die Auflösung für diesen Bereich (d.h. die Unterscheidbarkeit zwischen den ausgewählten Bakterientypen) verbessert werden.

Verschiedene Messmethoden priorisieren Sequenzanzahl, Fehlerrate und Sequenzlänge unterschiedlich. Bei aktuell sinkenden Kosten und methodischen Verbesserungen werden die Anzahl der Messungen und die Länge der sequenzierten Sequenzen tendenziell immer größer. Die Sequenzdaten ermitteln sich meist aus der Konstruktion eines neuen DNA-Strangs und können durch Qualitätskontrollen und Vergleiche mit Datenbanken modifiziert werden. Der bioinformatische Arbeitsablauf lässt sich durch modular-aufgebaute Open-Source Programme (Pipelines oder Frameworks) benutzerfreundlich anpassen und durchführen. Hierbei wird momentan insbesondere "qiime" [Caporaso *et al.* (2010)] benutzt.

Wenn einige Verzerrungen minimiert und vernachlässigt werden, hat die DNA jedes (beispielsweise) Bakterientyps ähnliche Wahrscheinlichkeit sequenziert zu werden und die gemessenen empirischen Bakterien-Verteilung entsprechen den ökologischen Verhältnissen. Aus verschiedenen Gründen werden die Verzerrungen zur Zeit nur schlecht minimiert (siehe Anhang C.2.5, Seite 163). Für die Theorie wird dies weitgehend vernachlässigt.

Genauere Informationen zum Next Generation Sequencing befinden sich in Anhang C, Seite 154.

2.3. Phylogenetische Verwandtschaft

Bei zwei ähnlichen Organismen, kann die Anzahl Sequenzunterschiede, die sie trennen, als ungefähres Zeitmaß betrachtet werden, welches die Zeitspanne, die zwischen ihnen und ihrem letzten gemeinsamen Vorfahren liegt, angibt. Wenn Sequenzabschnitte gewählt werden, in denen kein Austausch von DNA-Sequenzen stattgefunden hat und die Mutationshäufigkeit ungefähr der Erwartung entspricht, lässt sich eine Baumstruktur rekonstruieren, die die Verwandtschaftsstruktur der Organismen wiedergibt.

Mit zunehmender Zeitspanne wird dieses Modell schlechter. Trivialerweise steigen auch die Wahrscheinlichkeiten dafür, dass ein Basenpaar mehrfach mutiert oder dieselbe Mutation in zwei kaum verwandten Organismen stattfindet. Bäume können dann, auch wegen der oben erwähnten komplexeren Genomumstrukturierungen, nur noch aufwendig geschätzt werden und direkte Sequenzvergleiche eignen sich noch schlechter dazu, Verwandtschaftsbeziehungen zu erklären.

Abstände, die die Verwandtschaftsbeziehung zwischen Organismen widerspiegeln sollen, werden daher meist nicht direkt aus Sequenzvergleichen berechnet, sondern über den Umweg der Konstruktion eines phylogenetisches Baumes. Dies sind Bäume, die die Verwandtschaftsbeziehung genau beschreiben sollen. Bei der Konstruktion der Bäume geht Information aller Sequenzen mit ein. Hierdurch kann an einigen Stellen geklärt werden, ob Mutationen

z.B. mehrfach entstanden sind oder ob sie für eine gemeinsame Abstammung sprechen. Z.B. kann das Verwandtschaftsverhältnis zu anderen Organismen klarer aus den Sequenzen hervorgehen als zu dem gerade interessierenden Paar. Die Einordnung des Paars kann dann auf Umwegen über den Vergleich mit anderen Organismen zustande kommen.

Häufig wird sich in der Präsentation auf taxonomische Bäume beschränkt, die, im Gegensatz zu phylogenetischen Bäumen, stufenartig aufgebaut sind. Die geordneten Stufen sind: Königreich, Stamm, Klasse, Ordnung, Familie, Gattung und Art. Alle Kategorien der Organismen einer Stufe werden auf derselben Höhe der Y-Achse gezeichnet.

2.4. Pilzsequenzen

Für Pilze werden üblicherweise keine phylogenetischen Analysen aus den Sequenzierungsdaten erstellt, weil die traditionell verwendeten Erkennungssequenzen für Pilze hierfür besonders schlecht geeignet sind – siehe Anhang C.2.1.1, Seite 160. Stattdessen werden diese Sequenzen benutzt, um die Pilze in Datenbanken zu identifizieren. Wenn der Pilz bekannt ist und seine Verwandtschaftsbeziehungen schon einmal durch andere Sequenzen bestimmt wurden, können diese möglicherweise aus einer Datenbank bezogen werden. In Next Generation Sequencing Experimenten treten zur Zeit allerdings viele Organismen auf, die in noch keiner Datenbank zu finden sind.

Bedauerlich ist die Tatsache, dass bei der Standardisierung der für Pilze verwendeten Erkennungssequenz auch eine Sequenz zur Auswahl stand, die den fehlenden phylogenetischen Zusammenhang und weitere Probleme für Mikrobiomstudien nicht aufweist, sich aber das "Consortium for the Barcode of Life" im letzten Schritt des Auswahlprozess gegen diese entschieden hat – siehe Anhang C.2.1.2, Seite C.2.1.2. Andererseits gibt es sowieso noch weitere Probleme bei Pilzen: Diese Organismen sind wesentlich komplizierter als Bakterien. Sie besitzen mehrere Chromosome, können mehrere Kerne in einer Zelle haben und auch aus mehreren Zellen bestehen. Im Gegensatz zu Bakterien sind die bioinformatischen Methoden für Pilze viel weniger weit entwickelt. Für Grundlagenforschung vorverarbeiteter Mikrobiomdaten eignen sich Bakterien besser. Vom Anwendungsstandpunkt sind aber manchmal gerade Pilzdaten für Next Generation Sequencing interessant. So sind bei weitem nicht alle human-pathogenen Pilze kultivierbar und Pilzinfektionen sind generell ohne Next Generation Sequencing schlecht früh diagnostizierbar [Tiew *et al.* (2020)].

2.5. Verwendete Daten

Der Großteil der Arbeit basiert allerdings auf Bodenpilzen. Daten (Pilze und Bakterien) zu Versuchen an Nutzpflanzen stammen von Kooperationspartnern aus dem Julius Kühn-Institut in Braunschweig. Weitere Daten wurden von der Kosmetikfirma s-Biomedic zur Verfügung gestellt. Bei diesen Messungen handelt es sich um Bakterienpopulationen auf der Haut des

Gesichts von Aknepatienten. Den Patienten wurden Bakterien transplantiert, die mit einer Akne-freien Haut assoziert werden, in der Hoffnung, dass diese erstens angenommen werden und sich dauerhaft auf dem betroffen Patienten ansiedeln und zweitens tatsächlich eine positive Wirkung ausüben. Letzteres wäre der erste Wirksamkeitsnachweis eines "probiotischen" Konsumerprodukts. Dies konnte allerdings erst in einem Folgeversuch [Paetzold *et al.* (2019)] erfolgreich gezeigt werden, da die vorliegenden Daten ohne Plazebogruppe erhoben wurden.

Ein weiterer Datensatz von Bakteriengemeinschaften aus Bodenproben wurde von Neilson *et al.* (2017) erhoben. Auch dieser wurde verwendet. Er wird momentan von qiime als Beispiel verwendet. Das hatte den Vorteil, dass eine mögliche Fehlerquelle ausgeschlossen werden kann, da die bioinformatische Prozessierung schrittweise nach Vorgabe abgearbeitet werden konnte. Außerdem war zu erwarten, dass ein Datensatz, der keine Zusammenhänge zwischen Phylogenie und Versuchsdesign aufweist, nicht als Beispiel ausgewählt werden würde.

Zusätzlich zu realen Daten wurden auch Simulationen verwendet. Da der Zusammenhang zwischen Phylogenie und relativer Mikrobenhäufigkeit schwer abzusehen ist, wurde viel Wert auf eine umfassende Literaturrecherche der beteiligten Strukturen und möglicher Verbindungen gelegt.

2.6. Fortschritte während der Arbeit

Während der Anfertigung dieser Dissertationsarbeit gab es in allen Bereichen deutliche Fortschritte. Fortschritte in der bioinformatischen Verarbeitung haben dazugeführt, dass moderne Clusterungen von ausgelesenen Sequenzen nun Fehlertendenzen der Sequenziermethoden berücksichtigen und nicht mehr als OTUs bezeichnet werden, sonders einfach als "Sequenzen" [Callahan *et al.* (2016)]. Das einflussreiche bioinformatische Framework "qiime", bietet die entsprechenden Einstellungen seit dem Versionssprung zu "qiime 2" im Januar 2018 [Bolyen *et al.* (2019)] als Standardeinstellung an und verwendet und empfiehlt sie in Tutorien und in Kursen. Dort wird allerdings die Benennung "Feature" vorgeschlagen, um dem Anwender eine abstraktere Bedienungsoberfläche bieten zu können, die auch nichtsequenzielle Daten abdeckt. In neuen Ergebnissen wird dementsprechend kaum noch von OTUs gesprochen werden. Die Ergebnisse, die dieser Dissertation zugrunde liegen, wurden jedoch vorher aufbearbeitet, ausgewertet und veröffentlicht. Daher wurde auch ein zusätzlicher Datensatz untersucht, der mit modernen Methoden bestimmt wurde. Theoretisch hat sich an den grundlegendsten Prinzipien der Daten wenig geändert. Die Daten und ihre innere Struktur wurden jedoch wesentlich genauer. Die Anzahl an Variablen in einem typischen Datensatz kann um eine Größenordnung schrumpfen, da viele von den Variablen, die nur selten gemessen wurden, nun genauer zugeordnet und als ungenaue Sequenzmessungen einer anderen Variablen identifiziert werden. In neueren Daten könnten allerdings strukturelle Zusammenhänge sichtbar werden, die bei der OTU-basierten Auswertung verdeckt waren. Hier wird weiterhin von OTUs gesprochen werden, weil die meisten untersuchten Datensätze diese enthalten. Nur für den neuen Datensatz wird davon abgewichen.

3. Statistische Grundlagen

Hypothesentests dienen dazu, die Evidenz für potentielle Entscheidungen unter statistischen Gesichtspunkten zu bewerten.

Definition 3.1 (Hypothesentests). [Shao (2003)]
Sei $\mathcal{H}$ eine Menge von Wahrscheinlichkeitsmaßen und $H_0 \cup H_1 = \mathcal{H}, H_0 \cap H_1 = \emptyset$ eine Partitionierung von $\mathcal{H}$. H_0 wird Null-Hypothese genannt und H_1 Alternativ-Hypothese.
Sei $H \in \mathcal{H}$ und sei $X \sim H$ ein Zufallsvektor.
Jede Statistik Φ von dem Wertebereich von X nach $\{0, 1\}$ heißt (nicht-randomisierter) Test. Im Fall $\Phi(X) = 1$ sagt man, dass die Null-Hypothese H_0 abgelehnt wird und im Fall $\Phi(X) = 0$, dass sie akzeptiert wird.
Wird H_0 abgelehnt, obwohl $H \in H_0$ ist, spricht man von einem Fehler erster Art.
Wird H_0 akzeptiert, obwohl $H \in H_1$ ist, spricht man von einem Fehler zweiter Art.

Für die Wahrscheinlichkeit eines Fehlers erster Art kann eine obere Schranke α festgelegt werden $\sup_{H \in H_0}(P_H(\Phi(X) = 1)) \leq \alpha$, die Niveau des Tests genannt wird.
Ein Test $\Phi = \Phi_\alpha$ kann mittels einer Statistik T und einer Konstante $c_\alpha \in \mathbf{R}$ definiert werden:

$$T(x) \leq c_\alpha \Leftrightarrow \Phi(x) = 0$$
$$T(x) > c_\alpha \Leftrightarrow \Phi(x) = 1$$

α hängt dann monoton mit c_α zusammen. Für eine Realisierung x von X heißt in so einem Fall $p = \inf_{\substack{\alpha \in [0,1] \\ \Phi_\alpha(x)=1}} (\alpha)$ "p-Wert".

Bemerkung 1:
Den Begriff "Hypothese" werde ich synonym mit dem Begriff "Null-Hypothese" benutzen und "Alternative" synonym mit "Alternativ-Hypothese". In Shao (2003) ist der Niveaubegriff nicht auf die Null-Hypothese festgelegt, sondern kann auch für die Alternative vereinbart werden. Hier wird er allerdings ausschließlich für die Null-Hypothese verwendet.

3.1. Multiples Testen

Bei der gemeinsamen Betrachtung mehrerer zu prüfender Hypothesen steigt mit jeder wahren Hypothese, die zu einem Niveau größer 0 getestet wird, die Wahrscheinlichkeit, dass mindestens ein Fehler 1. Art auftritt. Diese Wahrscheinlichkeit wird dann familienweise Fehlerrate (FWER) genannt. Um diese zu beschränken, sind mehrere Verfahren bekannt, von denen einige hier vorgestellt werden.

3.1.1. Bonferroni-Korrektur

Seien $p_1, \ldots, p_k$ die p-Werte von $k \in \mathbf{N}_{>0}$ Tests. Seien weiterhin $b_1, \ldots, b_k \geq 1$ reelle Zahlen mit $\frac{1}{b_1} + \ldots + \frac{1}{b_k} = 1$ Größen zum gewichten der einzelnen Hypothesen. Dann sind die zugehörigen Bonferroni-adjustierten p-Werte: $\tilde{p}_1 := \min(b_1 p_1, 1), \ldots, \tilde{p}_k := \min(b_k p_k, 1)$. Für den Fall, dass $b_1 = \ldots = b_k = k$ gilt, wird das Verfahren auch "ungewichtete Bonferroni-Adjustierung" genannt und, falls die Gleichung nicht gilt, "gewichtete". [Shaffer (1995)]

3.1.2. Fixed-Sequence-Hierarchical-Tests

Beim Testen in einer vorbestimmten Reihenfolge können p-Werte, die weiter hinten stehen, durch die p-Werte davor adjustiert werden. Seien $p_1, \ldots, p_m$ für das Verfahren geeignet angeordnete p-Werte zu $m \in \mathbf{N}_{>0}$ Tests, die ihr Niveau einhalten. Definiere für jedes $i \in \{1, \ldots, m\}$ die adjustierten p-Werte $\tilde{p}_i := \max_{j \leq i}(p_j)$. Die so adjustierten p-Werte halten die FWER ein. Es gibt mehrere Möglichkeiten eine Reihenfolge zu finden, die geeignet ist. Eine Möglichkeit besteht darin, die Reihenfolge festzulegen, bevor die den Tests zugrundeliegenden Daten bekannt sind. Hier beschränkt der erste Test mit wahrer Hypothese die Wahrscheinlichkeit für einen Fehler erster Art in mindestens einem Test. [Qiu (2014)]

Datenabhängig geordnete Verfahren: Seien X ein m-dimensionaler multivariat-normalverteilter Zufallsvektor und $H_1, \ldots, H_m$ die univariaten Null-Hypothesen für die entsprechenden Dimensionen. Nach Kropf (2000) ergibt sich bei gewöhnlichen linearen Modellen eine geeignete Reihenfolge für die eben beschriebene Fixed-Sequence-Hierarchical Prozedur, wenn sie durch eine Permutation σ gegeben ist, für die gilt, dass Var $X_{\sigma(1)} \geq \ldots \geq$ Var$X_{\sigma(m)}$ ist. Diese Totalen-Varianzen bilden sich aus der Streuung innerhalb von Gruppen und den Verschiebungen zwischen ihnen. Wenn die Tests also nach monoton fallender Varianz angeordnet werden, ist, bei annähernd gleich-skalierten Variablen, zu erwarten, dass Variablen mit falscher Hypothese tendenziell weiter vorne in der Reihenfolge stehen. Das gilt besonders für Variablen mit großem Effekt, die häufig am relevantesten für die Praxis sind.

In Abschnitt 8.1 werden Mikrobiomdaten so strukturiert, dass Voraussetzungen vorliegen, die eine hohe Güte des Verfahrens begünstigen.

3.1.3. Abschlusstest-Prinzip

Für Betrachtungen bereits leicht komplexerer Anordnungen von Hypothesen sind folgende Notationen und Definitionen hilfreich:

Notation 3. Das Symbol $\subset$ bezeichne die Teilmengenrelation. Das Symbol $\subsetneq$ gilt bei echten Untermengen: $\big(a \subset b\big) = \big(a \subsetneq b \vee a = b\big)$, $\big(a \subsetneq b\big) = \big(a \subset b \wedge a \neq b\big)$. Die leere Menge wird bezeichnet durch $\emptyset := \{\}$. Die Anzahl der Elemente einer Menge M werde durch $\#M$ bezeichnet.

Definition 3.2. Bezeichne $\mathcal{P}(M)$ die Potenzmenge einer Menge M.
Für $m \in \mathbf{N}_{>0}$ seien $H_1, \ldots, H_m$ Hypothesen.
Für jedes $q \in \mathcal{P}(\{1, \ldots, m\}) \backslash \{\emptyset\}$ wird $\bigcap_{i \in q} H_i$ "Schnitthypothese" genannt.

Für alle $\emptyset \neq r \subset q$ gilt, dass die Ablehnung von $\bigcap_{i \in r} H_i$ die Ablehnung von $\bigcap_{i \in q} H_q$ impliziert. Dies motiviert ein Adjustierungsschema auf dieser Hypothesenmenge.

Satz 3.1 (Abschlusstestprinzip):
Seien für alle $q \in \mathcal{P}(\{1, \ldots, m\}) \backslash \{\emptyset\}$ die p-Werte p_q zu den Schnitthypothesen gegeben. Dann sind $\tilde{p}_q := \max_{r \supset q}(p_r)$ die durch das Abschlusstest-Prinzip adjustierten p-Werte.

Sei q die Menge aller Indices zu wahren Hypothesen und $r \in \mathcal{P}(\{1, \ldots, m\})$ eine Menge für die gilt $\exists_{i \in r} : H_i$ ist wahr. Dann ist $\tilde{p}_r \geq p_q$.
Wenn der Test zu p_q die Fehlerwahrscheinlichkeit eingehalten hat, halten die adjustierten p-Werte somit die familienweise Fehlerwahrscheinlichkeit ein. [Shaffer (1995)]

Der Index q ist hier eine Menge, funktioniert aber genauso wie eine gewöhnliche Zahl.

3.1.4. Hierarchisches Testen auf Bäumen

Für hierarchische Testverfahren empfiehlt es sich, die allgemeine Definition B.1 eines Baums etwas einzuschränken.

Definition 3.3. Für ein $m \in \mathbf{N}_{>0}$ sei $Q, \emptyset \neq Q \neq \{\emptyset\}$, eine Teilmenge der Potenzmenge $\mathcal{P}(\{1,\ldots,m\})$. Es gelte $\{1\},\ldots,\{m\},\{1,\ldots,m\} \in Q$.
Die ein-elementigen Mengen in Q werden "Blätter" und die m-elementige "Wurzel" genannt.
Es gelte die hierarchische Bedingung:

$$\forall_{q,r \in Q} : q \subset r \vee r \subset q \vee q \cap r = \emptyset \quad .$$

Für $q, r \in Q$, $q \subsetneq r$ wird q Nachfahre von r und r Vorfahre von q genannt. Ein Nachfahre q von r heißt Kind von r, wenn es keinen Vorfahren von q gibt, der Nachfahre von r ist. Wenn q Kind von r ist, wird r Vater von q genannt. Jedes $q \in Q$ ist ein Knoten.

Sei $k \in \mathbf{N}_{>0}$. Mengen der Struktur $Q_1,\ldots,Q_k \subset Q = Q_1 \cup,\ldots,\cup Q_k$ definieren die "Ebenen" einer hierarchischen Baumstruktur, mit der Wurzelebene $\{\{1,\ldots,m\}\} =: Q_k$ und der Blattebene $\{\{1\},\ldots,\{m\}\} =: Q_1$, falls für alle $i < k$ die Bedingungen gelten:

- $q \in Q_i \Rightarrow \exists_{r \in Q_{i+1}} : r$ ist Vater von q .
- $q, r \in Q_i \Rightarrow \Big(q = r \vee q \cap r = \emptyset\Big)$.

Eine Ebene mit höherem Index liegt "höher" als eine Ebene mit niedrigerem Index. Zu gegebenen Baum lässt sich die Ebene Q_i auch als "Ebene i" bezeichnen.

Eine Hierarchie werde "vollständig" genannt, wenn für alle $1 \leq i \leq k$ gilt, dass $\bigcup_{q \in Q_i} q = \{1,\ldots,m\}$.

Beim hierarchischen Testen nach Meinshausen (2008) wird im allgemeinen nur eine Teilmenge aller Schnitthypothesen betrachtet und adjustiert.

Satz 3.2 (Hierarchisches Testen nach Meinshausen (2008)):
$H_1,\ldots,H_m$ seien $m \in \mathbf{N}_{>0}$ Hypothesen. Q sei ein geeigneter Baum mit der Struktur von Definition 3.3, hierarchisch vollständig oder unvollständig. Für jedes $q \in Q$ wird die Schnitthypothese $H_q := \bigcap_{i \in q} H_i$ betrachtet, mit unadjustiertem p-Wert p_q.
Dann halten die folgend-adjustierten p-Werte $\tilde{p}'_q := \max_{r \in Q, q \subset r}(\tilde{p}_r)$ mit $\tilde{p}_q := \min\left(1, p_q \frac{m}{\#q}\right)$ die FWER ein.

Die Baumstruktur ist beispielsweise geeignet, wenn sie ohne Kenntnis der Messdaten definiert wurde.

3.2. Test auf Äquivalenz

3.2.1. Univariat

Wenn zwei unabhängige Größen X und Y mit Hilfe einer Statistik T auf „Gleichheit“ überprüft werden sollen, handelt es sich um die Fragestellung eines Äquivalenztests. Für diese Art der Problembetrachtung stellen sich zwei grundlegende Probleme:

In Äquivalenztests existieren in der Regel keine Punktnullhypothesen, weswegen viele Ansätze, die bei Tests auf Unterschied verwendet werden, nicht ohne weiteres angewendet werden können. Insbesondere betrifft dies Permutationstests.

Das zweite Problem liegt an der Definition der „Gleichheit“. In Tests auf Unterschied kann die Einschätzung über eine praktisch relevante Ungleichheit in einem zweiten Schritt geschehen: Nach dem Nachweis der Signifikanz kann die Relevanz der Ungleichheit anhand einer Effektgröße eingeschätzt werden. Im univariaten Fall lassen sich beide Betrachtungen gleichzeitig durch ein gewöhnliches Konfidenzintervall anstellen.

Für Tests auf Äquivalenz bietet es sich an, Zufälligkeit und Relevanz nicht getrennt, sondern zusammen zu betrachten. Falls die Relevanz durch zwei Schranken a und b angegeben werden kann, wird statt der Hypothese $E(T(X,Y)) \neq 0$, bezüglich einer Teststatistik T, die Hypothese $H : E(T(X,Y)) < a \vee E(T(X,Y)) > b$ verwendet. Ein Test zum Niveau α ergibt sich durch die Frage, ob ein $(1-2\alpha)$-Konfidenzintervall $C(X,Y)$ zu $E(T(X,Y))$ komplett in $[a,b]$ enthalten ist: $C(X,Y) \subset [a,b]$. Wenn das Intervall vollständig in $[a,b]$ liegt, ist mit Niveau α eine Äquivalenz nachgewiesen worden. Das ungewöhnliche Niveau des Konfidenzintervals erklärt sich aus einer Zerlegung der zusammengesetzten Hypothese H in die zwei einzelnen Hypothesen $H_1 : E(T(X,Y)) < a$ und $H_2 : E(T(X,Y)) > b$, die beide zum Niveau α abgelehnt werden müssen, damit der Test als signifikant angesehen wird.

3.2.2. Multivariat

Im multivariaten Fall mit m-Dimensionen lassen sich alle m Variablen einzeln betrachten, womit das Problem auf m univariate Fälle zurückgeführt werden kann. Wenn alle m univariaten Äquivalenz-Hypothesen zum Niveau α abgelehnt werden, ist die Äquivalenz auch für den multivariaten Fall zum Niveau α nachgewiesen. Anstatt m univariate Hypothesen zu betrachten, kann auch eine einzelne Hypothese formuliert und direkt getestet werden. Eine Methode, die auf paarweisen Abständen beruht, wird in Abschnitt 4.2 ausgearbeitet.

Bemerkung 2:
In Verfahren, die auf Linearkombinationen der Endpunkte aufbauen, besteht das Problem der unterschiedlichen inhaltlichen Gewichtung der Endpunkte – gelegentlich provokativ dargestellt als: Score = x mal Tod + y mal Schuhgröße. Bei abstandsbasierten Verfahren ist es die Aufgabe der Abstände, die Variablen geeignet zu bewerten. In einigen Fällen existieren gute oder brauchbare Abstände, die bereits in der Praxis angewandt werden. In anderen Fällen ist mitunter ein

multiples Verfahren vorzuziehen. Aufgrund des, dem Äquivalenzgedankens zugrunde liegenden IUT-Gedanken („Intersection-Union"), siehe Berger (1982), können die Endpunkte auch auf beide Ansätze aufgeteilt werden, ohne dass das Testniveau adjustiert werden muss. Auf diese Weise können unterschiedliche Endpunkte unterschiedlich oder unterschiedlich streng beachtet werden. Die Power fällt hingegen in der Regel, da beide Tests signifikant sein müssen, damit dies auch für den Gesamttest gilt.

3.3. Randomisierungstests

Randomisierungstests (Edgington (1995)) basieren auf der Hypothese, dass die Zuordnung der Versuchsteilnehmer in ihre Gruppen (bei unverbundenen Stichproben) oder der individuellen Messtagesumstände in ihre Reihenfolge (bei verbundenen Stichproben) einer permutationsinvarianten Verteilung entspricht, d.h. für n Gruppen-Zuordnungen G mit Realisation g gilt die Hypothese:

$$\forall_{\sigma \text{ ist Permutation von } \{1,\ldots,n\}} : P(G = g = (g_1, \ldots, g_n)) = P(G = (g_{\sigma(1)}, \ldots, g_{\sigma(n)})) \,.$$

Es wird dementsprechend keine Aussage über Grundgesamtheiten von Messwerten getroffen. Randomisierungstests setzen kein „random sampling“ voraus, sondern nur ein „random assignment“. Selbst kleine Stichproben sind oft signifikant und exakt, da die Tests nur Aussagen über Effekte in den jeweils vorliegenden Stichproben machen. Der Rückschluss auf eine Grund(-Teil-)population ist entkoppelt von dem Testentscheid und wird durch fachliche Beurteilung des Anwenders zur Repräsentativität der Stichprobe getroffen.

Bemerkung 3:
Der Einwand, dass Randomisierungstests nicht exakt seien, da der p-Wert nur in diskreten Abstufungen möglich ist, stimmt in soweit, dass in einem konkreten Randomisierungsdesign nur endlich viele Wahrscheinlichkeiten für einen Fehler erster Art möglich sind. Fordert man ein Niveau wie 5% ohne auf die tatsächliche möglichen Fehlerwahrscheinlichkeiten zu achten, dann führt das Standardvorgehen eines Vergleichs des p-Werts gegen das vorgegebene Niveau zu einem strikt konservativen Test, wenn sich das Niveau nicht unter den möglichen Fehlerwahrscheinlickeiten befindet.

Bemerkung 4:
Die Alternative eines Randomisierungstests ist immer das Komplement der Hypothese, also die Menge aller Verteilungen der Teststatistik, die nicht durch gleichverteilt-zufällige Zuordnung der vorliegenden Stichprobe entstehen. Sollen Aussagen über Verschiebungen des Testparameters getroffen werden, sind Randomisierungstests in der Regel nur approximativ und eventuell liberal.

Randomisierungstests lassen sich auf Abstandsstatistiken komplexer ökologischer Daten anwenden. Insbesondere können multivariate Tests problemlos auf phylogenetischen Abständen aufgebaut werden. Sie lassen sich leider schlecht sowohl für Äquivalenztests als auch für multiple Testverfahren der einzelnen Messwertdimensionen anwenden.

Die Hypothese eines Randomisierungstests hängt nicht mit dem Messwert zusammen, sondern nur mit der Gruppeneinteilung. Die ist für jede Messwertdimension identisch. Daher ist die Hypothese für jede Variable gleichzeitig wahr oder gleichzeitig falsch. Man kann Überlegungen darüber anstellen, dass die Randomisierungstest (abgesehen von der Hypothese zur Gruppeneinteilung) auch Hypothesen über die Verteilungen einzelner Variablen testen könnten. In diesen Fällen wird nicht von Randomisierungstests gesprochen (Edgington (1995)), sondern die allgemeinere Formulierung “Permutationstest” benutzt. Das wurde hier aber nicht weiter verfolgt.

Äquivalenztests gehen von der Nullhypothese aus, dass sich Gruppen unterscheiden, während die Gruppengleichheit unter der Nullhypothese wichtig für die Anwendbarkeit von Randomisierungstests ist.

Obwohl ökologische Probleme sich normalerweise elegant durch Randomisierungstests betrachten lassen, gilt das nicht für die beiden Teilbereiche, mit denen sich dieser Text befasst.

Teil II.

Äquivalenztest

Überblick über die Kapitel zu Äquivalenztests

In diesem Teil werden einige Grundlagen für einen Äquivalenztest für Mikrobiomdaten zusammengestellt. Die Tests sollen auf Konfidenzintervallen von Statistiken paarweiser Abstandsmaße aufbauen, die bereits in Abschnitt 3.2 besprochen wurden.

- In Abschnitt 4.1 wird eine Methode beschrieben, mit der sich Äquivalenzgrenzen bestimmen lassen.
- In Abschnitt 4.2 werden paarweise Abstandsmaße zu einer Statistik zusammengesetzt.
- In Abschnitt 4.3 werden Verfahren zur Berechnung von Konfidenzintervallen zu dieser Statistik vorgestellt.
- Ein auch für kleine Stichproben aus Experimenten mit praktisch gut realisierbaren Designs geeignetes Verfahren zur Varianzbestimmung, wird in Kapitel 5 hergeleitet.
- In Kapitel 7 wird die Eignung verschiedener Konfidenzintervalle durch Simulationen betrachtet.

Zusammen ist dadurch ein geeignetes Verfahren zur Überprüfung der Gleichheit von Mikrobengemeinschaften etabliert. Mehrere Anwendungen auf reale Datensätze finden sich in Antweiler *et al.* (2017).

4. Grundlagen für Äquivalenztests

4.1. Versuchsgetriebene Grenzenbestimmung

Der Nachweis der Gleichheit einer Behandlungsmethode zu einer Kontrollgruppe in einem oder mehreren Endpunkten erfordert eine festgelegte Grenze für den Unterschied ab dem zwei Populationen nicht mehr als "gleich" gelten sollen. Diese Grenze ist ein nicht-trivialer Bestandteil des Tests. In Abschnitt 4.1 wird die Bestimmung dieser Grenze betrachtet.

Bemerkung 5:
Bei Tests auf Unterschied hat sich die Ansicht durchgesetzt, dass in der Regel eine Gruppe mehr notwendig ist, die Kontrollgruppe, wenn der Anwender nicht in der Lage ist, ausreichende Angaben über den Prozess ohne den experimentellen Einfluss, aber mit Plazebo-Effekt zu treffen. Für Äquivalenztests hat sich eine entsprechende Ansicht, bezogen auf die Fähigkeit des Anwenders Aussagen über die Äquivalenzgrenze zu treffen und der daraus resultierenden Konsequenz, eine weitere Gruppe mit in die Analyse aufzunehmen, bisher nicht durchgesetzt.

4.1.1. Experimentell gegen Grundgesamtheit

Ein nicht-relevanter Unterschied wird bei diesem Vorgehen geschätzt, indem man den Unterschied zwischen einer Gruppe unter geeignetem experimentellen Einfluss und einer Kontrollgruppe schätzt. Es muss angenommen werden, dass dieser Unterschied irrelevant ist. In Abbildung 4.1 entspricht dies den blau dargestellten Elementen. Von statistischer Seite kann die jeweils konservative Seite (im Sinne des anschließenden Test) des Konfidenzintervalls des Unterschieds als Wert für die Grenze gefordert werden. Mit der bestimmten Grenze lässt sich eine Fallzahlplanung für den eigentlichen Test durchführen. Mit dieser zweiten Stichprobe wird schließlich der Test durchgeführt. In Abbildung 4.1 ist dies grün dargestellt. Die Bestimmung der Grenzen kann sowohl im selben als auch in einem eigenen Experiment durchgeführt werden und auch aus Daten eines bereits durchgeführten geeigneten Experiments ermittelt werden.

Beispiel 1. Die Verabreichungsform eines Medikaments soll geändert werden. Es ist bekannt, dass das Medikament zu unterschiedlichen Tageszeiten unterschiedlich gut aufgenommen wird. Es wird aber angenommen, dass diese Unterschiede für dieses Medikament nicht so relevant sind, dass der Anwendungszeitpunkt vorgeschrieben werden

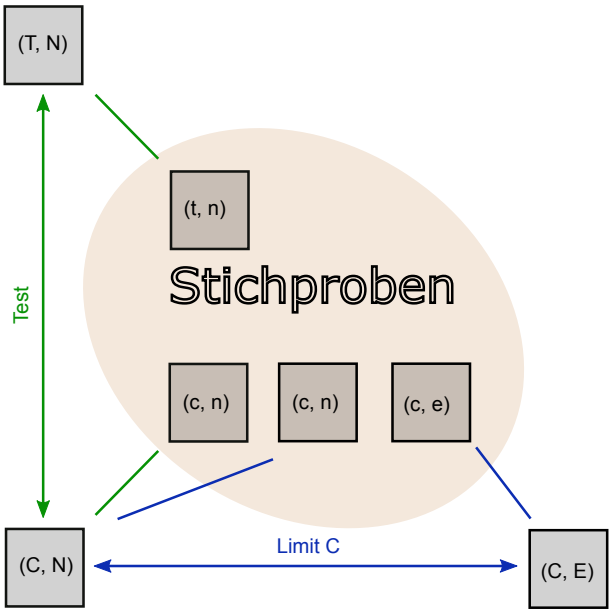

Abbildung 4.1.: Versuchsgetrieben optimal

(„,): Verschiedene Merkmale einer Gruppe
GROß/klein-Buchstaben: Grundgesamtheit/Stichprobe
C/T: Kontroll/Behandlungs-Gruppe
N/E: keine/eine zusätzliche experimentelle Beeinflussung
$\leftrightarrow$: Vergleiche
— : Schätzungen
grün: Für Test-Konfidenzintervall
blau: Für Äquivalenzgrenze

müsste - unter anderem, weil es seit Jahren in Umlauf ist und es keinen relevanten Hinweis darauf gegeben hat, dass er praktisch relevant sei.

Eine repräsentative Gruppe von Probanden enthält nun das Medikament mal morgens und mal abends. Die Konzentrationen werden an einer geeigneten Stelle im Körper gemessen. Der gemessene Unterschied der Konzentrationen wird als Grenze für den Unterschied im eigentlichen Test verwendet.

Ob der Mittelwert der Unterschiede genommen wird oder das Minimum oder der Unterschied von besonders geeigneten Probanden, unterliegt einer fachlichen Beurteilung der Situation. Hier sei der Maximalwert genommen worden, nachdem einige Probanden für die Bestimmung ausgeschlossen wurden, weil der Versuchsablauf bei ihnen nicht optimal funktioniert hat.

Das Medikament mit seiner veränderten Abreichungsform wird nun für den eigentlichen Test an eine Gruppe von Probanden gegeben. Für den Unterschied der Konzentrationen zur ursprünglichen Dareichungsform, wird ein Konfidenzintervall berechnet. Das Konfidenzintervall für den Unterschied liegt unterhalb der experimentell bestimm-

ten Grenze. Die Hypothese der Nicht-Äquivalenz wird abgelehnt.

Um mit weniger Probanden auskommen zu können, hätte ein anderer experimenteller Einfluss mit stärkerem aber trotzdem noch vernachlässigbarem Effekt verwendet werden können. Es sei bekannt, dass das Medikament zusammen mit Tee eingenommen, schlechter aufgenommen werde, aber dieser Effekt vernachlässigbar ist. Die Ermittlung der Grenzen führe zu höheren Werten, wodurch der eigentliche Test an Güte gewinnt.

Hätte die Gleichwertigkeit auf individueller Ebene bestimmt werden sollen, wäre ein experimenteller Einfluss für eine weitere Grenze notwendig gewesen. Beispielsweise könnte das Medikament zusammen mit und ohne Tee eingenommen werden, wobei dann für jeden einzelnen Probanden überprüft wird, ob bei ihm zwischen den beiden Einnahmen die Grenze aus den Tageszeiten überschritten wird. Der Anteil der Überschreitungen kann dann als Grenze für die Anzahl der Probanden mit Überschreitung in dem eigentlichen Versuch genommen werden.

In einigen Fällen ist eine getrennte Betrachtung für Ober- und Untergrenzen sinnvoll. Hier sind gegebenenfalls unterschiedliche experimentelle Einflüsse für die Bestimmung der beiden Grenzen notwendig.

Besonders relevant ist die experimentelle Bestimmung aber für komplizierte Größen – wie in Antweiler *et al.* (2017).

4.1.2. Pragmatisch

Im biologischen Forschungsbereich ist zur Zeit eine Fallzahlplanung in der Regel nicht üblich, da die realistisch bewerkstelligbaren Anzahlen zu klein sind. Daher bietet es sich an, die Grenzen erst während des Versuchs selbst zu bestimmen. Von dem Standpunkt aus betrachtet, dass nicht der Unterschied zwischen abstrakten Grundgesamtheiten zur Definition der Grenzen hinzugezogen wird, sondern der Unterschied zwischen zwei Gruppen mit nichtpathologischer Realisierung verwendet wird, kann eine Stichprobengruppe sowohl für die Grenzenbestimmung als auch für den Test verwendet werden (Bild 4.2).

Bemerkung 6:
Die Situation sei hier nochmals formaler ausgedrückt. Für drei unabhängige Zufallsvektoren X, Y, Z, von denen X und Y ohne experimentellen Zusatzeinfluss zustandekommen und Z mit, und X und Z aus den Kontrollbedingungen und Y aus den Behandlungsbedingungen, gilt für eine Teststatistik T unter der Nullhypothese: $P\big(T(X,Y) \leq k\big) \leq \alpha$. Wobei die Äquivalenzgrenze k eine feste aber unbekannte Zahl und α das Niveau des Tests sein soll. Für k gelte nach Einschätzung des Anwenders: $T(X,Z) \leq k$. Seien x und z Realisierungen der entsprechenden

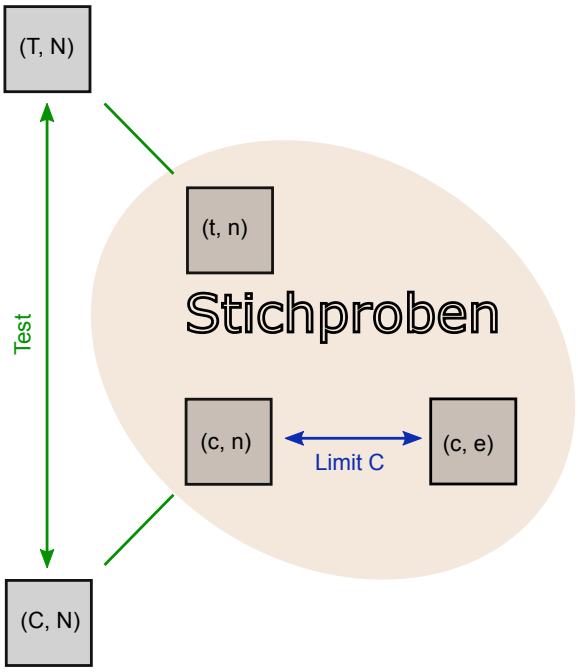

Abbildung 4.2.: Realisationsbasiert

(.,.): Verschiedene Merkmale einer Gruppe
GROß/klein-Buchstaben: Grundgesamtheit/Stichprobe
C/T: Kontroll/Behandlungs-Gruppe
N/E: keine/eine zusätzliche experimentelle Beeinflussung
$\leftrightarrow$: Vergleiche
$-$: Schätzungen
grün: Für Test-Konfidenzintervall
blau: Für Äquivalenzgrenze

Zufallsvektoren. Wenn für den Test die zum Durchführungszeitpunkt bekannte Größe $T(x,z)$ anstatt des unbekannten k verwendet wird, hält der Test sein Niveau weiterhin ein, da

$$P\big(T(X,Y) \leq T(X,Z)\big) = P\big(T(X,Y) \leq T(X,Z) \leq k\big) \leq P\big(T(X,Y) \leq k\big) \leq \alpha\,.$$

In einem vollen faktoriellen Design bietet es sich an, beide Gruppen, die zur Grenzenbestimmung benutzt wurden in den Test mit aufzunehmen und den Äquivalenztest nach dem zusätzlichen experimentellen Einfluss stratifiziert durchzuführen.

Bemerkung 7:
Die Situation sei hier nochmals formaler ausgedrückt: Seien X, Y, Z, z, k, T und α wie in Bemerkung 6 und W ein von X, Y und Z unabhängiger Zufallsvektor unter Behandlungsbedingung und mit experimentellem Zusatzeinfluss.
Unter der Nullhypothese gelte: $P\big(T(X,Y) \leq k,\ T(Z,W) \leq k\big) \leq \alpha\,.$

Wenn für den Test, die zum Durchführungszeitpunkt bekannte Größe $T(x,z)$ anstatt des unbekannten k verwendet wird, hält der Test sein Niveau weiterhin ein, da

$$\begin{aligned}
&P\big(T(X,Y) \leq T(X,Z),\ T(Z,W) \leq T(X,Z)\big)\\
=&P\big(T(X,Y) \leq T(X,Z) \leq k,\ T(Z,W) \leq T(X,Z) \leq k\big)\\
\leq&P\big(T(X,Y) \leq k,\ T(Z,W) \leq k\big) \leq \alpha\,.
\end{aligned}$$

4.2. Äquivalenztests auf Basis von paarweisen Abstandsmaßen

Spätestens nachdem ein Experiment für die Bestimmung der Äquivalenzgrenzen durchgeführt wurde, muss definiert werden, wie aus den Messdaten eine testbare Statistik berechnet werden soll. Die Statistik kann auf der Grundlage von Abstandsmaßen konstruiert werden. Durch paarweise Abstandsmaße lassen sich multivariate Probleme auf einen besonderen eindimensionalen Fall zurückführen. Das Resultat eines Abstands ist eine nicht-negative Zahl, mit der ein oberes Konfidenzintervallende für den Abstand zwischen den Gruppen bestimmt wird. Die Wahl des Abstandsmaßes hat großen Einfluss auf die resultierende Verteilung und deren Übereinstimmung mit theoretischen Annahmen. In Anhang A werden Betrachtungen und Simulationen in dieser Hinsicht durchgeführt. Einige Abstände, wie die euklidische Metrik, können aus Normen konstruiert werden. Aber auch kompliziertere Konstruktionen sind möglich und werden verwendet.

4.2.1. Ökologische Abstände

Relative Häufigkeiten von OTUs können als eine empirische Verteilung angesehen werden (Abb. 4.3). Zu diesen lassen sich dann Kenngrößen wie ihre Entropie bestimmen.

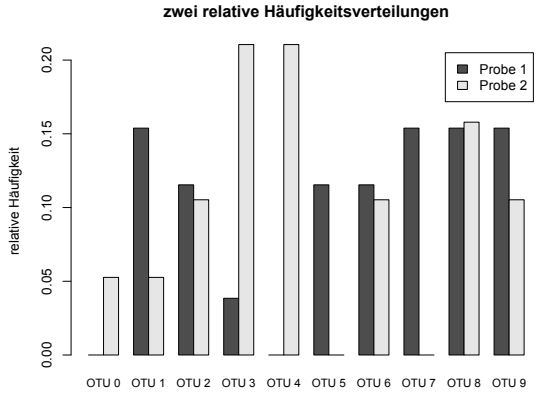

Abbildung 4.3.: Häufigkeitsverteilungen:

Zwei Proben mit jeweils einer relativen Häufigkeitsverteilung von 10 OTUs. Die OTUs 0 und 4 kommen in Probe 1 nicht vor und die OTUs 5 und 7 fehlen in Probe 2. Diese Verteilungen wurden zu Anschauungszwecken simuliert. Reale Daten enthalten erheblich mehr OTUs, der Anteil fehlender OTUs ist wesentlich größer und die Balkenhöhen sind viel ungleicher.

In der Agrarwissenschaft liegt für jedes Stichprobenelement (Pflanze), eine empirische Wahrscheinlichkeitsverteilung der OTUs der Mikrobenpopulation ihrer Wurzelregion vor. Unterschiede zwischen den Mikrobenpopulationen zweier Pflanzen entsprechen demnach Abständen zwischen zwei Verteilungen (Abb. 4.4).

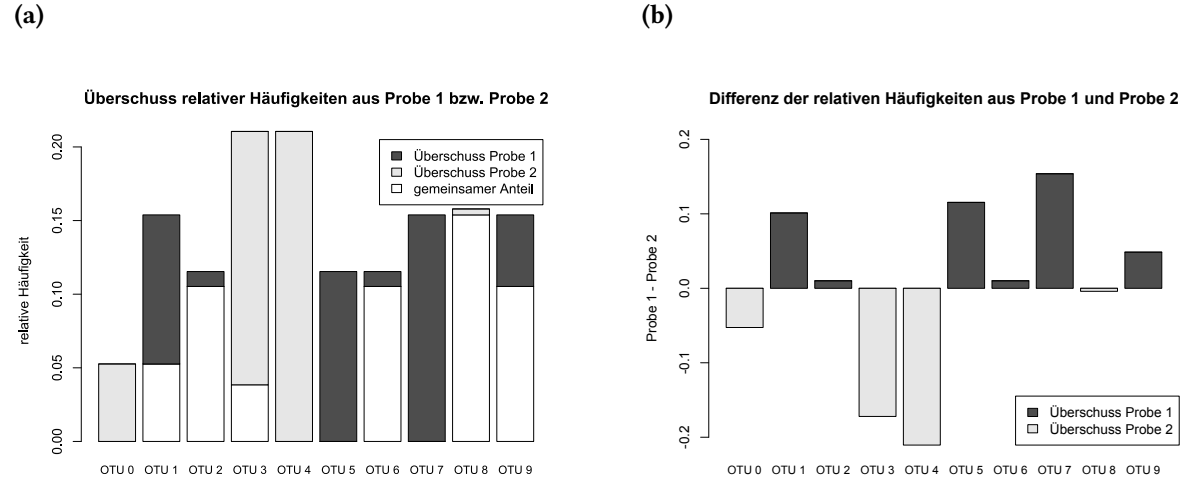

Abbildung 4.4.: Verteilungsabstand

In den Abbildungen ist für jede OTU der Dichteanteil dargestellt, den eine Probe mehr hat als die andere. Dabei handelt es sich um die simulierten Proben aus Abbildung 4.3. Dieser Anteil wird in den Abbildungen a und b "Überschuss" genannt.
Abbildung (a) zeigt die Verteilungen der beiden Proben übereinandergelegt. Jeder Balken reicht daher bis zum größeren Wert der entsprechenden OTU von den beiden Proben. Der kleinere Wert entspricht dem gemeinsamen Anteil dieser OTU. Der Balkenbereich darunter ist weiß gezeichnet und der darüber hat die Farbe der Probe, die den höheren Wert hat.
Abbildung (b) zeigt die Differenz der beiden Proben. Da die Y-Achse, verglichen mit Abbildung (a), die doppelte Spannweite bei identischem Flächenbedarf abdeckt, wirken die Balken halb so groß. Die Hälfte der Summe der Beträge dieser relativen Häufigkeitsdifferenzen definiert den relativen Bray-Curtis Abstand. Dieser wird in der Ökologie als Abstand zwischen den empirischen Verteilungen von Lebewesen aus zwei Habitaten verwendet.

Im ökologischen Kontext werden Abstände in der Regel "Betadiversität" genannt.

"Betadiversität" misst die Diversität zwischen Gemeinschaften. Soll phylogenetische Information mitberücksichtigt werden, wird meist der UniFrac Abstand verwendet. Für Analysen ohne Berücksichtigung phylogenetischer Information wird meist der Bray-Curtis- oder der Canberra-Abstand verwendet. Der ungewichtete UniFrac unterscheidet sich vom gewichteten UniFrac Abstand dadurch, dass present/absent Daten anstatt Häufigkeiten benutzt werden. [Knight *et al.* (2018)]

Der Bray-Curtis Abstand wurde in dieser Arbeit für die Konstruktion von Äquivalenztests aufgrund seiner Beliebtheit und seiner Eigenschaften bevorzugt.

Definition 4.1 (Bray-Curtis Abstand). [De Caceres *et al.* (2013)]
Für n OTUs seinen, $a_1, \ldots, a_n$ die Häufigkeiten der jeweiligen OTU in der ersten Probe und $b_1, \ldots, b_n$ die in der zweiten. Dann ist der Bray-Curtis Abstand zwischen den beiden Proben definiert als:

$$1 - \frac{2\sum_{i=1}^{n} \min(a_i, b_i)}{\sum_{i=1}^{n} a_i + \sum_{i=1}^{n} b_i} .$$

Bemerkung 8:
Für relative Häufigkeiten ist der Bray-Curtis Abstand offensichtlich gleich dem Abstand der halben 1-Norm, denn:

$$2 = \sum_{i=1}^{n} a_i + b_i = \sum_{i=1}^{n} \max(a_i, b_i) + \min(a_i, b_i)$$

und somit:

$$\begin{aligned}
&1 - \frac{2\sum_{i=1}^{n} \min(a_i, b_i)}{1+1} \\
=&\frac{1}{2}\left(\sum_{i=1}^{n} \max(a_i, b_i) + \min(a_i, b_i)\right) - \sum_{i=1}^{n} \min(a_i, b_i) \\
=&\frac{1}{2}\left(\sum_{i=1}^{n} \max(a_i, b_i) - \min(a_i, b_i))\right) \\
=&\frac{1}{2}\sum_{i=1}^{n} |a_i - b_i| = \frac{1}{2}||a - b||_1 \ .
\end{aligned}$$

Definition 4.2 (ungewichteter UniFrac Abstand). [Lozupone & Knight (2005)]
Gegeben sind zwei Stichproben und ein phylogenetischer Baum, der genau alle OTUs beider Proben als Blätter enthält. Die Kantengewichte bilden die evolutionäre Verschiedenheit der verbundenen Knoten ab und werden auch als Längen der Kanten bezeichnet. Alle Kanten, die sowohl auf einem Ast zu einer OTU der ersten als auch auf einem Ast zu einer OTU der zweiten Probe liegen, werden entfernt. Die Summe der Gewichte der verbleibenden Kanten ist der UniFrac-Abstand.

Anstatt nur die Anwesenheit von OTUs in den Proben zu berücksichtigen, können auch deren relative Häufigkeiten zur Skalierung der Kanten verwendet und dazu Differenzen gebildet werden. Darauf basiert der gewichtete UniFrac Abstand.

Definition 4.3 (gewichteter UniFrac Abstand). [Lozupone *et al.* (2007)]
Gegeben sei ein phylogenetischer Baum mit k Kanten. Für alle $1 \leq j \leq k$ sei $t_j \in \mathbf{R}_+$ die Länge der j-ten Kante. a_+ sei die Summe der Häufigkeiten aller OTUs der ersten Probe und b_+ die in der zweiten Probe. Für alle $1 \leq j \leq k$ sei A_j die Summe der Häufigkeiten der OTUs der ersten Probe, deren Ast von der Wurzel zu ihrem Blatt die Kante j enthält. Entsprechendes gelte für B_j und die zweite Probe. Der "rohe gewichtete UniFrac Abstand" ist definiert als:

$$u := \sum_{j=1}^{k} t_j \left| \frac{A_j}{a_+} - \frac{B_j}{b_+} \right| .$$

Der "normierte gewichtete UniFrac Abstand" wird gebildet, indem u noch durch einen Skalierungsfaktor D geteilt wird. Seien hierfür a_i, b_i und n wie in Definition 4.1.
Für alle $1 \leq i \leq n$ sei d_i die Länge des Asts von der Wurzel zum Blatt der OTU i.
Dann ist der "mittlere Abstand einer OTU zur Wurzel":

$$D := \sum_{i=1}^{n} d_i \left(\frac{a_i}{a_+} + \frac{b_i}{b_+} \right) .$$

Der normierte gewichtete UniFrac Abstand ist u/D.

Bemerkung 9:
Beide Versionen werden als "gewichteter UniFrac Abstand" bezeichnet und liefern in der Regel unterschiedliche Ergebnisse.

Weitere Information zu ökologischen Diversitätsmaßen befindet sich in Anhang B.3.11 auf Seite 152.

4.2.2. Abstände auf Stichproben

Abstandsmaße sind normalerweise nicht für ganze Stichproben definiert, sondern für einzelne Paare. Die resultierenden Abstandswerte für alle Paare der Stichprobe sind nicht mehr unabhängig und für viele statistische Verfahren nicht mehr direkt verwendbar. Diese Abstandswerte lassen sich, beispielsweise durch Konstruktion einer U-Statistik (Definition 5.1), für die ganze Stichprobe zusammenfassen, so dass eine einzige reelle Zahl für die gesamte Stichprobe angegeben werden kann. Diese kann sowohl für die Definition der Äquivalenzgrenzen als auch für die Teststatistik verwendet werden.

Beispielsweise kann für $X \sim N(\mu_x, \Sigma), Y \sim N(\mu_y, \Sigma)$ und ein beliebiges Abstandsmaß d durch den Mittelwert der paarweisen Abstände der Messwerte $X_1, \ldots, X_{n_x}$ aus X mit den

Messwerten $Y_1, \ldots, Y_{n_y}$ aus Y eine Statistik definiert werden:

$$r_b := \frac{1}{n_x n_y} \sum_{i=1}^{n_x} \sum_{j=1}^{n_y} d(X_i, Y_j)\,. \tag{4.1}$$

Interessiert man sich für den Abstand der beiden Zentroide, dann benötigt man bei diesem Ansatz einen Korrekturterm, der den Einfluss der Abstände innerhalb einer jeden Gruppe korrigiert. Wenn der mittlere Abstand innerhalb der Gruppe $X_1, \ldots, X_{n_x}$ bzw. $Y_1, \ldots, Y_{n_y}$, $n_x, n_y \in \mathbf{N}_{>0}$ gegeben ist durch:

$$r_X := \frac{1}{{n_x}^2} \sum_{i=1}^{n_x} \sum_{j=1}^{n_x} d(X_i, X_j)\,, \quad r_Y := \frac{1}{{n_y}^2} \sum_{i=1}^{n_y} \sum_{j=1}^{n_y} d(Y_i, Y_j)\,, \tag{4.2}$$

bietet sich, nach Satz A.1 auf Seite 101, falls d das Quadrat des euklidischen Abstands ist, folgende Statistik an:

$$S := r_b - \frac{r_X + r_Y}{2}\,. \tag{4.3}$$

4.2.2.1. Wahl der Übertragung eines Abstands auf Stichproben

Die Verteilung von $S := r_b$ aus Gleichung (4.1) ließ sich meist besser approximieren als die aus Gleichung (4.3) – siehe Abschnitt "normaler univariater euklidischer Abstand" auf Seite 103.

Das Abstandsmaß $S := r_b$ aus Gleichung (4.1) ist auch oft leichter zu interpretieren als S aus Gleichung (4.3), da der Mittelwert von Abständen zumindest immer eine reale Bedeutung hat, wohingegen Versuche, aus einzelnen Abständen die Abstände von Zentroiden zu konstruieren, nur in geeigneten Situationen gelingen. Wo es möglich ist, können stattdessen direkt die Zentroide in den Abstand eingesetzt werden. Nicht möglich ist das i.a. insbesondere, wenn der Abstand auch verzweigte Ausführungen (also "if"-Abfragen) zu kategoriellen Variablen enthält. Der ungewichtete UniFrac-Abstand, mit den present/absent-Daten, ist ein naheliegendes Beispiel, für das leicht zu erkennen ist, dass das Einsetzen von Mittelwerten machbar wäre, sich aber der Charakter des Abstandes ändern würde. Die Bedingung "Häufigkeit > 0" passt bei gemittelten Häufigkeiten nicht gut zum Charakter einer Mittelwertbildung.

Erwünschte Variabilität: Ein Herausrechnen von Innergruppen-Variabilität entfernt auch ein Unterscheidungsmerkmal, falls diese sich zwischen den Gruppen unterscheidet. Würde eine genetisch-homogene patentierte Pflanzenart in einer Gruppe mit natürlichen Äquivalenten verglichen werden, wäre es durchaus denkbar, dass jede der homogenen Pflanzen ein sehr ähnliches Bodenmikrobiom zeigenen würde, das auch in die natürliche Gruppe passen könnte, die Mikrobiome der natürlichen Gruppe aber untereinander eine hohe Variabilität aufweisen würden. Diese hohe Variabilität wäre ein ökologisch relevanter Unterschied zwischen den Gruppen. In einer solchen Situation wäre noch stärker darauf zu achten, dass

Variabilitäts-Unterschiede berücksichtigt werden, als es in Antweiler *et al.* (2017) getan wurde. Hierfür wäre ein zusätzlicher Äquivalenztest für die Absolut-Differenz $|r_X - r_Y|$ der mittleren Innergruppenabstände aus Gleichung (4.2) denkbar. Für unterschiedliche Kenngrößen können Messungen mit unterschiedlichen akzeptablen experimentellen Einflüssen erforderlich sein.

Mischung von Proben und Lebensräumen: Im ökologischen Kontext ist eine Mittelwertbildung vor Verwendung eines Abstandsmaßes als Vermischen der Proben unmittelbar vor der Analyse interpretierbar aber nicht unbedingt wünschenswert. Das Mischen von Proben kann zu nicht-repräsentativen Ergebnissen führen. In Fällen, in denen eine Gruppe Proben aus verschiedenen Lebensräumen enthält (z.B. trockener Sandboden und feuchter Lehmboden), kann der kombinierte Lebensraum andere Eigenschaften haben und somit über die Zeit hinweg andere Lebewesen selektieren als es bei einer Durchmischung unmittelbar vor der Analyse den Anschein hätte. Pflanzen und Bodentypen haben einen selektierenden Einfluss, während der Standort eher bestimmt, welche Mikroben überhaupt selektiert werden können (Anhang B.3.10.1, Seite 151). Selbst im Falle von identischen Lebensräumen, ist aufgrund von Stellenäquivalenz oder ungesättigten Gemeinschaften, damit zu rechnen, dass die gemischten Gemeinschaften nicht der zu untersuchenden Situation entsprechen (Anhang B.3.1, Seite 142). Ebensolche Effekte können durch unterschiedlichen Species-Turnover und früh angesiedelte Arten zustande kommen (Anhang B.3.2, Seite 144). Würde ein gemischter Lebensraum längere Zeit beobachtet, würde die höhere interspezifische Konkurrenz gemäß Konkurrenzausschlussprinzip zu gegenseitiger Beeinflussung der Häufigkeiten der Arten führen (Anhang B.3.1, Seite 142). Die interspezifische Konkurrenz in einer künstlich gemischten Probe wird daher unterschätzt. Unterschiedliche zeitliche Entwicklungsabschnitte (wie Regionen entlang einer wachsenden Wurzel) (Anhang B.3.10.1, Seite 151), zeitliche Abstände zu Störungen und unterschiedlichen Störungen (Anhang B.3.2, Seite 144) können bei Vermischung zu außergewöhnlichen Kombinationen von r- und K-Strategen (Anhang B.3.2, Seite 144) führen. Effekte aus unterschiedlichen Zeitpunkten eines Zyklus könnten hingegen in den Mischungen möglicherweise approximativ eher natürlichen Zuständen entsprechen (Anhang B.3.5, Seite 147).

Diversitätsmaße, wie die Entropie, reagieren in der Regel gewollt empfindlich auf Änderung der Artenvielfalt. Phylogenetische Information kann die Diskrepanz zur Realität durch künstliche oder physische Vermischung vermutlich etwas abschwächen, da nah-verwandte Arten häufig ähnliche Nischen haben und stärker miteinander konkurrieren (Anhang B.3.1, Seite 142), ihre Häufigkeits-Unterschiede aber von phylogenetischen Abständen häufig heruntergewichtet werden. Dass der Unterschied in diesen Maßen zwischen einzelnen und gemischten Proben angeglichen wird, mag ökologisch keine wünschenswerte Eigenschaft sein, führt aber zu mehr Freiheiten in der Auswertung.

Physische Vermischung von Proben findet auch in ökologischen Studien statt, wenn an verschiedenen Stellen um eine Pfanze Proben genommen und zu einer einzigen zusammen-

gefügt werden - oder auf kleinerer Ebene, wenn eine Probe bearbeitet wird, die verschiedene Kleinstlebensräume enthält. In den Situationen sind die gemischten Proben aber vergleichsweise ähnlich. Bei kleinen Fallzahlen kann eine Varianzreduktion notwendig sein und andere Nachteile überwiegen.

Für die Konstruktion des Äquivalenztests waren in Antweiler *et al.* (2017) die Annäherungen der Teststatistik an eine Normalverteilung ausschlaggebend.

4.3. Intervallschätzer

Die Ideen aus den vorherigen Abschnitten dieses Kapitels reichen aus, um Äquivalenzgrenzen für den Test zu berechnen, womit die erste Grundlage für einen Äquivalenztest erreicht ist. Einen Äquivalenztest auf parametrischen Verfahren für ökologische Maße auf Mikrobiomdaten aufzubauen erscheint zu kompliziert. Resamplingverfahren können die Test-Konstruktion vereinfachen, wobei es hinreichend ist, eine geeignete Methode zur Bestimmung eines Konfidenzintervalls zu einer Statistik S zu finden. Gemäß Abschnitt 3.2 lässt sich dann ein Äquivalenztest definieren. Einige der üblichen Verfahren zur Konstruktion von Konfidenzintervallen waren, insbesondere aufgrund der kleinen Stichprobengrößen, nicht geeignet.

4.3.1. Bootstrap

Als echte Bootstrap-Verfahren (Efron & Tibshirani (1993)) sind für den vorliegenden Text sowohl reines Resamplen auf den Stichprobendaten durch eine ungewichtete Multinomialverteilung als auch durch Schätzen von Mittelwert und Kovarianzmatrix der Stichprobe mit anschließenden Simulieren normalverteilter Daten mit diesen Parametern versucht worden.

Da stets mit geringen Fallzahlen gearbeitet wurde, waren keine guten Ergebnisse für die Überdeckungsrate des Konfidenzintervalls bezüglich der wahren Testgröße zu erwarten (Scholz (2007)) und sind auch nicht eingetreten. Rechnerisch-aufwendige Doppel-Bootstrap-Varianten, die in kleinen Stichproben besser funktionieren (Scholz (2007)), wurden hier nicht untersucht.

4.3.2. Jackknife

In dieser Arbeit wird das im folgenden Kapitel beschriebene 2-Stichproben-Jackknife-Verfahren verwendet um die Varianz einer Statistik S zu schätzen. Die Varianz wird dann zur Konstruktion eines Konfidenzintervalls $C(S)$ benutzt. Dies hielt das Niveau in Simulationsuntersuchungen deutlich besser ein als die anderen Verfahren und erscheint ausreichend zu sein in Szenarien, die dem späteren Versuch ähneln.

5. Jackknife

5.1. Zerlegung von Funktionen unabhängiger Zufallsvariablen

In diesem Abschnitt werden die Grundlagen bereitgestellt, mit denen sich komplexe Jackknife-Verfahren nicht-asymptotisch untersuchen lassen. Zu Notationszwecken wird unter anderem die Interpretierbarkeit von diskreten injektiven Inklusions-Abbildungen als Permutationen genutzt. Ebenso werden Multiindizes benutzt und durch Skalarprodukte charakterisiert — hierbei insbesondere die Charakterisierung als Unter-Multiindex eines anderen Multiindexes. Ein von Efron & Stein (1981) erweitertes (sie weisen allerdings darauf hin, dass andere Autoren Vergleichbares entwickelt haben) auf Hajek (1968) zurückgehendes Projektionsverfahren wird vorgestellt, mit dem sich Statistiken mit mehreren Zufallsargumenten systematisch zerlegen lassen. Dies ist von zentraler Bedeutung. Zunächst werden jedoch U-Statistiken eingeführt, für die sich später Aussagen über ihre Varianz in Abhängigkeit der Stichprobengröße treffen lassen.

Definition 5.1 (U-Statistik). [Hoeffding (1948)]
Seien $p, n, m \in \mathbf{N}_{>0}, n \geq m$ positive Zahlen,
$X_1, \ldots, X_n \in \mathbf{R}^p$ unabhängige Zufallsvektoren,
$f : \mathbf{R}^{pm} \to \mathbf{R}$ eine Funktion und
M die Menge aller injektiven Abbildungen $\pi : \{1, \ldots, m\} \to \{1, \ldots, n\}$,
dann nennt man $U = \sum_{\pi \in M} \frac{(n-m)!}{n!} f(X_{\pi(1)}, \ldots, X_{\pi(m)})$ eine U-Statistik.

Bemerkung 10:
M enthält $\binom{n}{m} m!$ Abbildungen. Falls f symmetrisch ist (in dem Sinne, dass die Reihenfolge der Argumente irrelevant ist), reicht es aus, über geeignete $\binom{n}{m}$ Abbildungen π zu mitteln, da durch die Permutationsinvarianz mindestens jeweils $m!$ Abbildungen $f \circ \pi$ übereinstimmen.

Beispiel 2. Seien $X_1, \ldots X_n$ i.i.d. verteilte Zufallsvariablen und $U := \frac{n}{n-1}(\overline{X^2} - \bar{X}^2)$ der erwartungstreue Varianzschätzer. Dann ist U eine U-Statistik, denn für

$f(a,b) := \frac{1}{2}(a-b)^2$ ist $m = 2$, $M = \{\pi : \{1,2\} \mapsto \{1,\dots,n\} | \pi(1) \neq \pi(2)\}$ und $\tilde{M} := \{\pi \in M | \pi(1) < \pi(2)\}$ ist eine geeignete Teilmenge von M. Somit ist:

$$\begin{aligned}
&\sum_{\pi\in\tilde{M}} \binom{n}{2}^{-1} f(X_{\pi(1)}, X_{\pi(2)}) = \sum_{1\le i<j\le n} \frac{1}{n(n-1)} (X_i - X_j)^2 \\
=&\sum_{i=1}^{n}\sum_{j=1}^{n} \frac{1}{2n(n-1)} (X_i - X_j)^2 = \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{1}{2n(n-1)} \left(2X_i^2 - 2X_iX_j\right) \\
=&\sum_{i=1}^{n} \frac{1}{n(n-1)} \left(nX_i^2 - nX_i\bar{X}\right) = \frac{n}{n-1}\left(\overline{X^2} - \bar{X}^2\right).
\end{aligned}$$

Definition 5.2 (Multiindex). Für $n \in \mathbf{N}$ ist ein Multiindex i definiert als ein n-Tupel $i \in \{0,1\}^n$. Für zwei Multiindizes i, j zum selben n gibt $i \wedge j$ das elementweise Produkt und $< i, j >$ das Skalarprodukt an.

Bemerkung 11:
$i \wedge j$ gibt die gemeinsamen Einsen von i und j an und das Skalarprodukt $< i, j >$ die Anzahl der gemeinsamen Einsen. Das Skalarprodukt von i mit sich selbst $< i, i >$ gibt dementsprechend die Anzahl der Einsen in i an. Unter-Multiindizes lassen sich unter anderem kompakt durch Skalarprodukte mit dem entsprechenden (Über-)Multiindex beschreiben. Für zwei Multiindizes i, j zum selben n gibt das Skalarprodukt in der Gleichung $< i, j - i >= 0$ bzw. $< i, j >=< i, i >$ an, dass i keine Einsen enthält, die j nicht enthält. Diese Eigenschaft definiert i zu einem Unter-Multiindex von j. Beide Darstellungen sind äquivalent. Da die erste etwas kürzer ist, wird sie häufiger in den Gleichungen dieses Kapitels verwendet als die intuitivere zweite Darstellung.

Um einen Multiindex zur Auswahl eines Zufallsvektors zu benutzen ist folgende Notation t_j oft hilfreich:

Notation 4.

$$\forall j \in \{1,\dots,n\}: \quad t_j : \begin{array}{c} \{0,1\}^n \to \{0,\dots,n\} \\ i \mapsto j * i_j \end{array}$$

Mit '$*$' sei hierbei die gewöhnliche Multiplikation in $\mathbf{N}_0$ bezeichnet.

Bemerkung 12:
$t_j(i)$ gibt die j-te Indexposition des Multiindex i an, falls dieser dort 1 ist. Andernfalls ist $t_j(i) = 0$.

Bemerkung 13:
Im weiteren wird $X_0 = 0$ konstant gewählt werden. t_j wird im Bedingungsteil bedingter Erwartungswerte verwendet werden. Die Auswahl von X_0 entspricht somit keiner zusätzlichen Bedingung.

Satz 5.1 (ANOVA-Typ-Zerlegung von Funktionen unabhängiger Zufallsvektoren):
Seien $n \in \mathbf{N}, X_0 := 0$ und $X_1, \ldots, X_n$ unabhängige Zufallsvektoren, und S eine Funktion derart, dass $S_n := S(X_1, \ldots, X_n)$ eine Zufallsvariable mit endlicher Varianz ist. Sei

$$\forall_{r \in \{0,1\}^n} A_r := \sum_{k=0}^{<r,r>} (-1)^{<r,r>-k} \sum_{\substack{j \in \{0,1\}^n \\ <r,j>=<j,j>=k}} E(S_n | X_{t_1(j)}, \ldots, X_{t_n(j)}) \,,$$

dann gilt: $S_n = \sum_{m=0}^{n} \sum_{\substack{r \in \{0,1\}^n \\ <r,r>=m}} A_r = \sum_{r \in \{0,1\}^n} A_r \,.$

Bemerkung 14:
A_r summiert mit alternierendem Vorzeichen die Summen bedingter Erwartungswerte von S_n, bezüglich der durch die Unter-Multiindizes des Multiindex r mit gleicher Anzahl an Einsen gegebenen Zufallsvektoren X_j. Es handelt sich somit um den Anteil von S_n, der auf genau die durch r definierten Zufallsvektoren zurückgeht, bereinigt um den Anteil, der sich durch echte Untermengen dieser Zufallsvektoren erklären lässt. Gezeigt wird, dass sich S_n in die Summe aller A_r zerlegen lässt.

Beweis. Seien $X_0, \ldots, X_n, S, S_n$ und A_r wie in Satz 5.1 gefordert.

$$
\begin{aligned}
\sum_{m=0}^{n} \sum_{\substack{r\in\{0,1\}^n \\ <r,r>=m}} A_r &= \sum_{m=0}^{n} \sum_{\substack{r\in\{0,1\}^n \\ <r,r>=m}} \sum_{k=0}^{m} (-1)^{m-k} \sum_{\substack{j\in\{0,1\}^n \\ <r,j>=<j,j>=k}} E(S_n|X_{t_1(j)},\ldots,X_{t_n(j)}) \\
&= \sum_{m=0}^{n}\sum_{k=0}^{m} (-1)^{m-k} \sum_{\substack{r\in\{0,1\}^n \\ <r,r>=m}} \sum_{\substack{j\in\{0,1\}^n \\ <r,j>=<j,j>=k}} E(S_n|X_{t_1(j)},\ldots,X_{t_n(j)}) \\
&= \sum_{m=0}^{n}\sum_{k=0}^{m} (-1)^{m-k} \binom{n-k}{m-k} \sum_{\substack{j\in\{0,1\}^n \\ <j,j>=k}} E(S_n|X_{t_1(j)},\ldots,X_{t_n(j)}) \qquad (5.1) \\
&= \sum_{k=0}^{n}\sum_{m=k}^{n} (-1)^{m-k} \binom{n-k}{m-k} \sum_{\substack{j\in\{0,1\}^n \\ <j,j>=k}} E(S_n|X_{t_1(j)},\ldots,X_{t_n(j)}) \\
&= \sum_{k=0}^{n}\sum_{m=0}^{n-k} (-1)^{m} \binom{n-k}{m} \sum_{\substack{j\in\{0,1\}^n \\ <j,j>=k}} E(S_n|X_{t_1(j)},\ldots,X_{t_n(j)}) \\
&= \sum_{\substack{j\in\{0,1\}^n \\ <j,j>=n}} E(S_n|X_{t_1(j)},\ldots,X_{t_n(j)}) = E(S_n|X_1,\ldots,X_n) = S_n
\end{aligned}
$$

Die letzte Zeile folgt aus:

$$
\sum_{m=0}^{n-k} 1^{n-k-m}(-1)^m \binom{n-k}{m} = \begin{cases} (1-1)^{n-k} = 0 \text{ , für } k \neq n \\ 1 \qquad\qquad\qquad \text{ , für } k = n \end{cases} .
$$

Gleichung (5.1) ergibt sich aus der vorherigen Zeile, aus der Beobachtung, dass alle Multiindizes j mit k Einsen gleich häufig in der Summe auftauchen. Diese Häufigkeit entspricht der Anzahl der Möglichkeiten aus den restlichen $n-k$ Indexpositionen $m-k$ auszuwählen, da k Einsen in der Zielgleichung bereits durch j bestimmt sind und nur die restlichen Einsen aus r variieren. □

Efron & Stein (1981) beweisen diesen Satz durch Berechnung der Koeffizienten der bedingten Erwartungswerte, führen dies allerdings nur für den Koeffizienten des auf nichts bedingenden Erwartungswerts aus. Karlin & Rinott (1982) beweisen den Satz ebenfalls durch Berechnung der Koeffizienten, beschreiben dies jedoch für alle Koeffizienten.

Beispiel 3. Seien $X_0 := 0, X_1 \sim N(1,1), X_2 \sim N(2,1)$ unabhängige Zufallsvariablen, $S_2 := S(X_1, X_2)$ und S die Funktion:

$$
\begin{aligned}
S : \mathbf{R} \times \mathbf{R} &\to \mathbf{R} \\
(a,b) &\mapsto ab + a + b + 3 \,.
\end{aligned}
$$

Die durch die vier möglichen Multiindizes r indizierten Terme A_r sind dann:

$$\begin{aligned}
A_{(0,0)} =&(-1)^{0-0}E(S_2|X_0,X_0) = E(S_2) = 1*2+1+2+3 = 8\,,\\
A_{(1,0)} =&(-1)^{1-0}E(S_2|X_0,X_0) + (-1)^{1-1}E(S_2|X_1,X_0)\\
=&E(S_2|X_1) - E(S_2) = (2X_1+X_1+2+3) - 8 = 3X_1 - 3\,,\\
A_{(0,1)} =&(-1)^{1-0}E(S_2|X_0,X_0) + (-1)^{1-1}E(S_2|X_0,X_2)\\
=&E(S_2|X_2) - E(S_2) = (X_2+1+X_2+3) - 8 = 2X_2 - 4\,,\\
A_{(1,1)} =&(-1)^{2-0}E(S_2|X_0,X_0) + (-1)^{2-1}E(S_2|X_1,X_0)\\
&+ (-1)^{2-1}E(S_2|X_0,X_2) + (-1)^{2-2}E(S_2|X_1,X_2)\\
=&E(S_2|X_1,X_2) - E(S_2|X_1) - E(S_2|X_2) + E(S_2)\\
=&(X_1X_2+X_1+X_2+3) - (2X_1+X_1+2+3) - (X_2+1+X_2+3) + 8\\
=&X_1X_2 - 2X_1 - X_2 + 2\,.
\end{aligned}$$

S_2 ergibt sich nun wieder aus:

$$\begin{aligned}
S_2 =&A_{(0,0)} + A_{(1,0)} + A_{(0,1)} + A_{(1,1)}\\
=&E(S_2) + (E(S_2|X_1) - E(S_2)) + (E(S_2|X_2) - E(S_2))\\
&+ (E(S_2|X_1,X_2) - E(S_2|X_1) - E(S_2|X_2) + E(S_2))\\
=&E(S_2|X_1,X_2)\,,
\end{aligned}$$

bzw. mit den ausgerechneten Termen der A_r :

$$\begin{aligned}
S_2 =&8 + (3X_1-3) + (2X_2-4) + (X_1X_2 - 2X_1 - X_2 + 2)\\
=&X_1X_2 + X_1 + X_2 + 3\,.
\end{aligned}$$

Für die Zerlegungsterme sind im Beispiel folgende Eigenschaften erkennbar, die im Weiteren allgemein bewiesen werden:

Zentriertheit bei positiven Indizes (Korollar 5.3, Seite 37)
Für alle $r \neq (0,0)$ ist $E(A_r) = 0$.

Bedingte Zentriertheit bei Indizes außerhalb der Bedingung (Lemma 5.2, Seite 36)
Für alle r mit $r_2 = 1$ ist $E(A_r|X_1) = 0$.
Für alle r mit $r_1 = 1$ ist $E(A_r|X_2) = 0$.

Paarweise Unkorreliertheit (Lemma 5.4, Seite 37)

$$\begin{aligned}
E(A_{(1,0)}A_{(0,1)}) =& E(6X_1X_2 - 12X_1 - 6X2 + 12) = 12 - 12 - 12 + 12 = 0 \\
E(A_{(1,1)}A_{(1,0)}) =& E(3X_1^2X_2 - 6X_1^2 - 3X_1X_2 + 6X_1 - 3X_1X_2 + 6X_1 + 3X_2 - 6) \\
=& 6E(X_1^2) - 6E(X_1^2) - 6 + 6 - 6 + 6 + 6 - 6 = 0 \\
E(A_{(1,1)}A_{(0,1)}) =& E(2X_1X_2^2 - 4X_1X_2 - 2X_2^2 + 4X_2 - 4X_1X_2 + 8X_1 + 4X_2 - 8) \\
=& 2E(X_2^2) - 8 - 2E(X_2^2) + 8 - 8 + 8 + 8 - 8 = 0
\end{aligned}$$

Für alle $r \neq (0,0)$ ist $E(A_r A_{(0,0)}) = 0$.

Notation 5. Zur einfacheren Notation seien im Folgenden:

$$\vec{t}(r) = t_1(r), \ldots, t_n(r) \qquad \text{und} \qquad X_{\vec{t}(r)} = X_{t_1(r)}, \ldots, X_{t_n(r)} \,.$$

Außerdem gelten die Bedingungen und Bezeichnungen von Satz 5.1 für den restlichen Abschnitt und die daran anschließenden Abhandlungen über Jackknife-Verfahren. Insbesondere gilt dies für X_i, S_n und A_r .

Für die Zerlegungsterme werden nun zwei Eigenschaften der Zentriertheit und Unkorreliertheit nachgewiesen.

Lemma 5.2 (Bedingte Zentriertheit bei Indizes außerhalb der Bedingung):
Seien r, q Multiindizes und es existiere ein $i \in \{1, \ldots, n\}$ mit $r_i = 1$ und $q_i = 0$, dann gilt:

$$E(A_r | X_{\vec{t}(q)}) = 0 \,.$$

Beweis. Seien r und q wie in Lemma 5.2 gefordert.

$$
\begin{aligned}
E(A_r|X_{\vec{t}(q)}) &= \sum_{k=0}^{n} \sum_{\substack{s\in\{0,1\}^n \\ <s,r>=<s,s>=k}} (-1)^{n-<r,s>} E(E(S_n|X_{\vec{t}(s)})|X_{\vec{t}(q)}) \\
&= \sum_{\substack{s\in\{0,1\}^n \\ <s,r>=<s,s>}} (-1)^{n-<r,s>} E(E(S_n|X_{\vec{t}(s)})|X_{\vec{t}(q)}) \\
&= \sum_{\substack{s\in\{0,1\}^n \\ <s,r>=<s,s>}} (-1)^{n-<r,s>} E(S_n|X_{\vec{t}(s\wedge q)}) \\
&= \sum_{\substack{s\in\{0,1\}^n \\ <s,r>=<s,s> \\ s_i=0}} (-1)^{n-<r,s>} E(S_n|X_{\vec{t}(s\wedge q)}) + \sum_{\substack{s\in\{0,1\}^n \\ <s,r>=<s,s> \\ s_i=1}} (-1)^{n-<r,s>} E(S_n|X_{\vec{t}(s\wedge q)}) \\
&= 0
\end{aligned}
$$

Die letzte Gleichung gilt, da i in $s \wedge q$ in beiden Fällen 0 indiziert, aber der Vorzeichenterm unterschiedlich ist, da $< s, r >$ von s_i abhängt. $\wedge$ ist hierbei das "bitweise und" d.h. eine elementweise Multiplikation in dem $\{0,1\}^n$ Raum. Der rechte Term existiert aufgrund der Voraussetzung $r_i = 1$. □

Efron & Stein (1981) sagen, dass der Beweis durch Berechnung der Koeffizienten durchführbar ist, ohne dies auszuführen. Karlin & Rinott (1982) erklären die hier verwendete Idee als Teil des Beweises zu Lemma 5.4.

Korollar 5.3 (Zentriertheit bei positiven Indizes):
Für alle $r \in \{0,1\}^n \backslash \{\vec{0}\}$ *gilt:* $E(A_r) = 0$.

Beweis. Nach Voraussetzung existiert ein i mit $r_i = 1$. Angewandt auf $E(A_r) = E(A_r|X_0)$ zeigt Lemma 5.2 die Aussage. □

Lemma 5.4 (Paarweise Unkorreliertheit):
Für alle Multiindizes $i, j, i \neq j$ *gilt:* $E(A_iA_j) = 0$.

Beweis. Seien $i, j \in \{0,1\}^n$ mit $i \neq j$. OBdA existiere ein l mit $i_l = 1$ und $j_l = 0$.

$$
E(A_iA_j) = EE(A_iA_j|X_{\vec{t}(j)}) = E(A_jE(A_i|X_{\vec{t}(j)})) \underset{(Lemma\ 5.2)}{=} 0
$$

□

Der Beweis entspricht genau dem von Karlin & Rinott (1982), ist hier allerdings über die Beweise von Lemma 5.2, Korollar 5.3 und Lemma 5.4 verteilt.

Bemerkung 15:
Lemma 5.4 gilt mit derselben Begründung auch dann, wenn A_i und A_j auf unterschiedlichen Zufallsvariablen $S(X_1, \ldots, X_n)$ und $S(Y_1, \ldots, Y_n)$ beruhen, wobei für alle $k \in \{1, \ldots, n\}$ gelten soll, dass $Y_k := X_k$ sein darf, aber alle anderen Zufallsvariablen von einander unabhängig sind. Hierbei ist die Einschränkung, dass identische Zufallsvariablen in S an derselben Stelle stehen müssen, auch nur relevant für die Notation. Für die Beweisidee und die Anwendungsmöglichkeiten ist dies nicht relevant.

Karlin & Rinott (1982) benutzen folgendes Hilfsmittel zur Analyse von Jackknife-Statistiken und beweisen die benötigten Eigenschaften dieses Hilfsmittel mittels der eben gezeigten Zerlegung.

Definition 5.3. Definiere für jeden beliebigen Multiindex r das generalisierte bedingte Varianzfunktional:

$$C_r := E\big((E(S_n|X_{\vec{t}(r)}))^2\big) = EE^2(S_n|X_{\vec{t}(r)})\,.$$

Notation 6. Seien $m \in \mathbf{N}_{>0}$, $\vec{r}, p \in \mathbf{N}^m$ und $\sum_{i=1}^m p_i = n$ mit $\forall_{i\in\{1,\ldots,m\}} r_i \leq p_i$.
Sei $S : (x_{11}, \ldots, x_{p_1 1}, \ldots, x_{1m}, \ldots, x_{p_m m}) \mapsto S(x_{11}, \ldots, x_{p_1 1}, \ldots, x_{1m}, \ldots, x_{p_m m})$ in den ersten p_1 Argumenten permutationsinvariant und ebenso in den nächsten p_2 und allen weiteren.
Seien die Zufallsvektoren $X_{11}, \ldots, X_{p_1 1}$ i.i.d. und ebenso alle weiteren Zufallsvektorenblöcke.
Seien hier, sowie im restlichen Kapitel, die Zufallsvektoren zusätzlich durchgängig indiziert: $X_1 := X_{11}, \ldots, X_{p_1} := X_{p_1 1}, X_{1+p_1} := X_{1,2}, \ldots, X_n := X_{p_m m}$. Sei $q \in \{0,1\}^n$. Dann ist nur die Anzahl der jeweiligen Argumente in einem Variablenblock für das verallgemeinerte bedingte Varianzfunktional C_q relevant, nicht aber ihre Anordnung. Unter diesen Bedingungen sei

$$c_{\vec{r}} := C_q \qquad \text{mit:} \qquad \forall_{i\in\{1,\ldots,m\}} \quad r_i := \sum_{j=1}^{p_i} q_{j+\sum_{k=1}^{i-1} p_k}\,.$$

Bemerkung 16:
Die i.i.d.-Bedingung wird durch die Bedingung der Permutations-Invarianz in den Blöcken nötig für $c_{\vec{r}}$ und wird auch in den Jackknife-Schätzern gefordert werden. Für C_r und Satz 5.5 ist sie nicht nötig.

Beispiel 4. Seien $X_0 := 0, X_1 := X_{11}, X_2 := X_{21} \sim N(1,1), X_3 := X_{12} \sim N(2,1)$ unabhängige Zufallsvariablen, $S_3 := S_{(2,1)} := S(X_{11}, X_{21}, X_{12})$ und S die Funktion:

$$\begin{aligned} S : \mathbf{R}^2 \times \mathbf{R} &\to \mathbf{R} \\ (x_{11}, x_{21}, x_{12}) &\mapsto x_{11}x_{21} + x_{12} - 2\,. \end{aligned}$$

Dann sind

$$\begin{aligned} c_{(2,0)} =& C_{(1,1,0)} \\ =& EE^2(X_{11}X_{21} + X_{12} - 2|X_{11}, X_{21}) \\ =& E((X_{11}X_{21} + 2 - 2)^2) = E(X_{11}^2)E(X_{21}^2) = 4 \\ & \text{denn: } E(X_{11}^2) = \mathrm{Var}(X_{11}) + E^2X_{11} = 1 + 1 \end{aligned}$$

und

$$\begin{aligned} c_{(1,0)} =& C_{(1,0,0)} = C_{(0,1,0)} \\ =& EE^2(X_{11}X_{21} + X_{12} - 2|X_{11}) \\ =& E((X_{11} + 2 - 2)^2) = E(X_{11}^2) = 2\,. \end{aligned}$$

$c_{\vec{r}}$ wird für die Jackknife-Betrachtungen von großer Bedeutung sein. Die nächste Bemerkung deutet bereits darauf hin.

Bemerkung 17:
Es gilt:

$$\begin{aligned} ES_n^2 &= c_{(p_1,\ldots,p_m)} = c_{\vec{p}} \\ E^2S_n &= c_{(0,\ldots,0)} = c_{\vec{0}} \\ & \textit{und somit:} \\ Var(S_n) &= c_{(p_1,\ldots,p_m)} - c_{(0,\ldots,0)} = c_{\vec{p}} - c_{\vec{0}}\,. \end{aligned}$$

Satz 5.5 (Zerlegung des generalisierten bedingten Varianzfunktionals):
Sei r ein Multiindex. Es gilt:

$$C_r = E\big(E(S_n|X_{\vec{l}(r)})^2\big) = \sum_{\substack{i\in\{0,1\}^n \\ <i-r,i>=0}} E\big(A_i^2\big)\,.$$

Beweis.

$$
\begin{aligned}
C_r =& E\Bigg(E\Bigg(\sum_{m=0}^{n}\sum_{\substack{q\in\{0,1\}^n\\ <q,q>=m}} A_q\Big|X_{\vec{t}(r)}\Bigg)^2\Bigg) = E\Bigg(E\Bigg(\sum_{q\in\{0,1\}^n} A_q\Big|X_{\vec{t}(r)}\Bigg)^2\Bigg)\\
=& E\Bigg(E\Bigg(\sum_{q\in\{0,1\}^n} A_q\Big|X_{\vec{t}(r)}\Bigg)E\Bigg(\sum_{s\in\{0,1\}^n} A_s\Big|X_{\vec{t}(r)}\Bigg)\Bigg)\\
=& \sum_{q\in\{0,1\}^n}\sum_{s\in\{0,1\}^n} E\big(E\big(A_q\big|X_{\vec{t}(r)}\big)E\big(A_s\big|X_{\vec{t}(r)}\big)\big)\\
\underset{(Lemma\ 5.2)}{=}& \sum_{\substack{q\in\{0,1\}^n\\ <q,q-r>=0}}\sum_{\substack{s\in\{0,1\}^n\\ <s,s-r>=0}} E\big(E\big(A_q\big|X_{\vec{t}(r)}\big)E\big(A_s\big|X_{\vec{t}(r)}\big)\big)\\
=& \sum_{\substack{q\in\{0,1\}^n\\ <q,q-r>=0}}\sum_{\substack{s\in\{0,1\}^n\\ <s,s-r>=0}} E\big(A_qA_s\big)\\
=& \sum_{\substack{q\in\{0,1\}^n\\ <q,q-r>=0}} E\Bigg(A_q^2+\sum_{\substack{s\in\{0,1\}^n\\ <s,s-r>=0\\ s\neq q}} A_qA_s\Bigg) \underset{(Lemma\ 5.4)}{=} \sum_{\substack{q\in\{0,1\}^n\\ <q,q-r>=0}} E(A_q^2)
\end{aligned}
$$

□

Die zweitletzte Zeile folgt daraus, dass q und s Untermultiindizes von r sind und somit alle Zufallsvariablen, die in A_s oder A_q eingehen, auch in der Bedingung stehen.

Dieser Beweis entspricht dem von Karlin & Rinott (1982).

Beispiel 5. Für die in Beispiel 4 angegeben Größen gilt:

$$
\begin{aligned}
E\big(A^2_{(0,0,0)}\big) =& E\Big(\big(E(X_{11}X_{21}+X_{12}-2)\big)^2\Big) = 1\,,\\
E\big(A^2_{(1,0,0)}\big) =& E\Big(\big(E(X_{11}X_{21}+X_{12}-2|X_{11})-A_{(0,0,0)}\big)^2\Big)\\
=& E\big((X_{11}-1)^2\big) = 1\\
=& E\big(A^2_{(0,1,0)}\big)\,,\\
E\big(A^2_{(1,1,0)}\big) =& E\Big(\big((X_{11}X_{21})-E(X_{11}X_{21}+X_{12}-2|X_{11})\\
& -E(X_{11}X_{21}+X_{12}-2|X_{21})+EA_{(0,0,0)}\big)^2\Big)\\
=& E\Big(\big(X_{11}X_{21}-X_{11}-X_{21}+1\big)^2\Big)\\
=& E\Big(\big(X_{11}X_{21}\big)^2-2X_{11}X_{21}(X_{11}+X_{21}-1)+\big(X_{11}+X_{21}-1\big)^2\Big)\\
=& 4-2(2+2-1)+(2+2+1+1+1-1-1-1-1)=1\,,
\end{aligned}
$$

$$
\begin{aligned}
C_{(1,1,0)} =& E\big(A^2_{(0,0,0)}\big)+E\big(A^2_{(1,0,0)}\big)+E\big(A^2_{(0,1,0)}\big)+E\big(A^2_{(1,1,0)}\big)\\
=& 1+1+1+1=4\\
=& c_{(2,0)}\\
=& \binom{2}{0}\binom{0}{0}E\big(A^2_{(0,0,0)}\big)+\binom{2}{1}\binom{0}{0}E\big(A^2_{(1,0,0)}\big)+\binom{2}{2}\binom{0}{0}E\big(A^2_{(1,1,0)}\big)\,.
\end{aligned}
$$

Korollar 5.6:
Für blockweise permutationsinvariante Funktionen mit i.i.d. Zufallsvektoren in jedem Argumentenblock lässt sich Satz 5.5 mittels Notation 6 offensichtlich vereinfachen zu:

$$c_{\vec{r}} = \sum_{i_1=0}^{p_1}\cdots\sum_{i_m=0}^{p_m}\prod_{j=1}^{m}\binom{r_j}{i_j}E(A^2_{\vec{i}})$$

mit $A_{\vec{i}} := A_\nu$ für einen beliebigen Multiindex ν mit:

$$\forall_{j\in\{1,\dots,m\}}\sum_{k=0}^{p_j}\nu_{k+\sum_{l=1}^{j-1}p_l} = i_j\,.$$

Für $d > b$ gelte per Definition $\binom{b}{d} = 0$.

Aus den jeweils r_j fixierten Zufallsvektoren werden i_j ausgewählt. Dafür gibt es $\binom{r_j}{i_j}$ Möglichkeiten. Die Summe über alle zulässigen i_j ergibt die Formel. Das "zulässig" entspricht der Unter-Multiindex-Eigenschaft. Die Bedingung für A_ν ist eine unhandliche Formulierung für die naheliegende Idee, dass ein geeigneter Multiindex gewählt werden kann.

5.2. Jackknife

Das Jackknife-Verfahren kann zur Reduktion des Bias eines Schätzers und zur Schätzung seiner Varianz benutzt werden. Es basiert auf Mittelungen mehrerer Schätzungen, bei denen stets nur ein Teil der zur Verfügung stehenden Stichprobenelemente benutzt wird. Stehen mehr Stichprobenelemente zur Verfügung als in den Schätzer eingesetzt werden können, dient das Verfahren auch zur Varianzreduktion. Auch in Fällen, in denen mehr Stichprobenelemente in den Schätzer eingesetzt werden können, kann es vorkommen, dass man sich auf eine geringere Anzahl beschränkt, um Verzerrungen zu vermeiden, die durch ungleiche Stichprobengrößen entstehen. In der Mikrobiomanalyse wird oft durch eine erzwungen-identische Anzahl von Reads versucht, einen Bias in Diversitätsschätzern zu vermeiden. Die Anzahl der Reads ist die Anzahl geglückter Sequenzierungen in einer Probe. Auch an dieser Stelle werden Jackknife-Verfahren verwendet, um alle Stichprobenelemente verwenden zu können, um die Varianz zu reduzieren.

Im Extremfall lassen sich zwei Stichproben auf jeweils ein Element einschränken. Auf diese Weise lassen sich paarweise Abstände auf naheliegende Weise über ein Jackknife-Verfahren verwenden. Eine andere Möglichkeit, paarweise Abstände zu benutzen, besteht darin, eine willkürliche Teststatistik aus allen paarweisen Abstandswerten zu erstellen und ein Jackknife-Verfahren zu benutzen, um die Varianz dieser Statistik zu schätzen.

Das Jackknife-Verfahren geht zurück auf Quenouille (1949), der es als Verfahren zur Bias-Reduktion vorschlug. Für Tests wurde es von Tukey (1958) vorgeschlagen und im weiteren auch als Varianzschätzer genutzt. Asymptotische Jackknife-Verfahren für Mehrgruppenprobleme wurden von Arvesen (1969) entwickelt. (Da die Stichprobengrößen in typischen biologischen Datensätzen klein sind, werden sie in dieser Arbeit nicht benutzt.) Nichtasymptotische Jackknife-Verfahren für Mehrgruppenprobleme mit geeigneten Eigenschaften wurden von Karlin & Rinott (1982) entwickelt.

Eine praxisrelevante Eigenschaft fehlt einem ihrer Verfahren. Das gruppenbezogene Verhältnis der Stichprobengröße und der zugehörigen Argumente der Funktion durfte sich nicht zwischen den Gruppen unterscheiden. Selbst bei geeigneten Statistiken und balanziert geplanten Designs kommt es mitunter vor, dass ein Stichprobenelement nicht zur Auswertung zur Verfügung steht. In der Agrarwissenschaft mag beispielsweise eine Pflanze nicht wachsen, die dementsprechend auch nicht sinnvoll weiter untersucht werden kann. Dann sind die Gruppengrößen nicht mehr so, wie sie geplant waren. Es ist wünschenswert, dass ein Verfahren keine strikten Ansprüche an die Gruppengrößen stellt. Ein weiteres Verfahren, das sie vorstellten, hatte diese Eigenschaft. Sie erkannten allerdings nicht, dass ihr anderes

Verfahren auch um diese Eigenschaft erweitert werden kann. Das Verfahren wird in diesem Abschnitt nun um die Eigenschaft erweitert.

Zunächst sei ein Einstichproben-Verfahren erwähnt, das weitgehend strukturgleich zu dem Mehrgruppen-Verfahren ist.

Satz 5.7 (Einstichproben Jackknife Varianzschätzer):
Seien $X_1, \ldots, X_n$ i.i.d. verteilte Zufallsvariablen, $p, k \in \mathbf{N}$ mit $1 \leq p < n = p + k$ und S eine permutationsinvariante Funktion derart, dass $S_p := S(X_1, \ldots, X_p)$ eine Zufallsvariable mit endlicher Varianz ist.
Sei I eine Menge aller $\binom{n}{k}$ bis auf Permutation unterschiedlichen möglichen Injektionen $\{1, \ldots, p\} \to \{1, \ldots, n\}$.
Sei $\bar{S}_k = \frac{1}{\binom{n}{k}} \sum_{i \in I} S_{k,i}$ der Mittelwert aller $\binom{n}{k}$ möglichen Zufallsvariablen $S_{k,i} := S(X_{i(1)}, \ldots, X_{i(p)})$. Ferner sei

$$V_k(S) := \frac{1}{\binom{n-1}{p}} \sum_{i \in I} (S_{k,i} - \bar{S}_k)^2 .$$

Dann gilt:

$$\mathit{Var}\, S_p(X_1, \ldots, X_p) \leq E\, V_k(S) .$$

Für die Zufallsvariable $T_p := \sum_{j=1}^{p} X_j$ gilt hier die Gleichheit.

Der Beweis ergibt sich unmittelbar aus Satz 5.8 mit $m = 1$.

Bemerkung 18:
Es ist auf den Unterschied zwischen den Bezeichnungen $\bar{S}_k$ und $S_{k,i}$ und der von S_p zu achten. Die ersten beiden sind Jackknife-spezifisch und geben die Anzahl zusätzlicher Elemente an. Das p in S_p besagt nur, dass die Zufallsvariable auf einer Funktion mit p Argumenten beruht.
V_k schätzt die Varianz der p-argumentigen Zufallsvariable S_p. Für den Fall, dass S_p als U-Statistik darstellbar ist, gilt folgende Ungleichung aus Hoeffding (1948):

$$p\ \mathit{Var}\, S_p \geq (p+1)\ \mathit{Var}\, S_{p+1} . \tag{5.2}$$

Die Varianz einer $p+1$ elementigen Form kann in diesem Fall nicht-liberal durch den Varianzschätzer der p elementigen Form abgeschätzt werden, selbst wenn diese Schätzung mit $\frac{p}{p+1}$ geschrumpft wird.

Bemerkung 19:
Das Einstichproben-Jackknife-Verfahren ist in seiner "leave one out"-Form, d.h. für den Fall $k = 1$, am bekanntesten, wobei der Dimensionskorrekturfaktor $(n-1)/n$ aus Gleichung (5.2) in der Regel direkt in den Formeln mitrepräsentiert wird, allerdings ohne auf Voraussetzungen für die Anwendung hinzuweisen.

Der Jackknifeschätzer aus Satz 5.7 lässt sich für mehrere Stichproben und oder Stratifizierungen verallgemeinern. Im folgenden Satz werden teilweise Bedingungen, für die eine endliche Menge ausreichen würde, für ganz $\mathbf{N}$ gefordert, da hieraus keine Einschränkungen entstehen und die ohnehin aufwendige Notation etwas übersichtlicher wird.

Satz 5.8 (Mehrgruppen-Jackknife):
Seien $X_{11}, \ldots, X_{n_1 1}$ *i.i.d.*; $X_{12}, \ldots, X_{n_2 2}$ *i.i.d.*; $\ldots$; $X_{1m}, \ldots, X_{n_m m}$ *i.i.d. unabhängige Zufallsvektoren, jeweils mit endlicher Varianz. Die Semikolons trennen die Zufallsvektoren in verschiedene Blöcke. Das "i.i.d." bezieht sich jeweils auf die Zufallsvektoren, die in dem entsprechenden Block stehen.*
Seien $0 < p_1, \ldots, p_m, k_1, \ldots, k_m \in \mathbf{N}_{>0}$ *mit* $\forall_{j \in \mathbf{N}} : p_j + k_j = n_j$ *und* $\gamma := \max_{j \in \mathbf{N}} p_j / n_j$.
Sei $S : (x_{11}, \ldots, x_{p_1 1}; \ldots; x_{1m}, \ldots, x_{p_m m}) \mapsto S(x_{11}, \ldots, x_{p_1 1}; \ldots; x_{1m}, \ldots, x_{p_m m})$ *eine Funktion, die invariant unter Permutation der ersten* n_1 *Argumente und ebenso unter den nächsten* n_2*, usw. bis zu den letzten* n_m *ist, mit der Eigenschaft, dass*
$S_{p_1,\ldots,p_m} := S(X_{11}, \ldots, X_{p_1 1}; \ldots; X_{1m}, \ldots, X_{p_m m})$ *eine Zufallsvariable mit endlicher Varianz ist.*
Für alle $j \in \mathbf{N}$ *sei* I_j *eine Menge aller* $\binom{n_j}{k_j}$ *bis auf Permutation unterschiedlichen möglichen Injektionen* $\{1, \ldots, p_j\} \to \{1, \ldots, n_j\}$.
Sei $\bar{S}_{k_1,\ldots,k_m} = \frac{1}{\prod_{j=1}^m \binom{n_j}{k_j}} \sum_{i_1 \in I_1, \ldots, i_m \in I_m} S_{k_1,\ldots,k_m,i_1,\ldots,i_m}$ *der Mittelwert aller* $\prod_{j=1}^m \binom{n_j}{k_j}$ *möglichen Zufallsvariablen der Form*
$S_{\vec{k},\vec{i}} := S_{k_1,\ldots,k_m,i_1,\ldots,i_m} := S(X_{i_1(1)1}, \ldots, X_{i_1(p_1)1}; \ldots; X_{i_m(1)m}, \ldots, X_{i_m(p_m)m})$.

Seien

$$K := K(p_1, \ldots, p_m, n_1, \ldots, n_m) := \frac{1}{(1-\gamma) \prod_{j=1}^m \binom{n_j}{p_j}}$$

und

$$V_{k_1,\ldots,k_m}(S) := K \sum_{i_1 \in I_1; \ldots; i_m \in I_m} (S_{k_1,\ldots,k_m,i_1,\ldots,i_m} - \bar{S}_{k_1,\ldots,k_m})^2 .$$

Dann gilt:

$$\mathit{Var}\, S_{p_1,\ldots,p_m}(X_{11}, \ldots, X_{p_1 1}; \ldots; X_{1m}, \ldots, X_{p_m m}) \leq EV_{k_1,\ldots,k_m}(S) . \tag{5.3}$$

Für die Zufallsvariable $T = X_{11} + \ldots + X_{p_1 1} + \ldots + X_{1m} + \ldots + X_{p_m m}$ *gilt hier die Gleichheit, falls* $\forall_{j \in \mathbf{N}} : p_j / n_j = \gamma$.

Bemerkung 20:
Satz 5.8 und der folgende Beweis sind Erweiterungen von denen, die Karlin & Rinott (1982) veröffentlicht haben. Der Satz, den sie "Theorem C" nennen, ist für den Fall $m = 2$ *angegeben, sie sagen jedoch, dass er auf beliebiges* m *verallgemeinerbar sei. Darüber hinaus unterscheidet er sich von Satz 5.8 darin, dass sie für den gesamten Satz fordern*

$$\exists_{\gamma \in \mathbf{Q}} \forall_{j \in \{1,\ldots,m\}} : p_j / n_j = \gamma .$$

Eine Verallgemeinerung, wie hier vorgestellt, hielten sie explizit für nicht möglich. Hier wird in dem Beweis zu Satz 5.8 gezeigt, dass eine Normalisierungskonstante gewählt werden kann, die für alle Statistiken, die die Voraussetzung des Satzes erfüllen, zum Einhalten der Varianzungleichung (5.3) *führt.*

Wie zu Beginn der Sektion beschrieben, macht diese Erweiterung das Verfahren für viele praktische Probleme erst nutzbar.

Der folgende Beweis beruht, genau wie Karlin und Rinotts, auf folgenden Prinzipien:

- *Die Ungleichung* (5.3) *wird dadurch nachgewiesen, dass gezeigt wird, dass die Differenz zwischen der Erwartung des Varianzschätzer und der Varianz der Zufallsvariable $S_{\vec{p}}$ nichtnegativ ist.*
- *Die Varianz wird mittels des generalisierten bedingten Varianzfunktionals in der Form aus Bemerkung 17 dargestellt.*
- *Der Varianzschätzer wird ebenfalls in generalisierte bedingte Varianzfunktionale zerlegt.*
- *Die Erwartungen der entstandenen Unterterme werden umgeordnet, mittels Korollar 5.6 ersetzt und mit den entsprechenden Termen, die aus der Varianz hervorgegangen sind, verrechnet.*
- *Dies ergibt eine Summe nichtnegativer Terme und zugehöriger Koeffizienten. Der Nachweis der Nichtnegativität dieser Koeffizienten schließt den Beweis von Ungleichung* (5.3) *ab.*

Den Nachweis, dass die Koeffizienten nichtnegativ sind, überlassen Karlin & Rinott (1982) ihren Lesern, mit dem Hinweis, dass dieser aufwendig sei. Die Gleichheit für lineare Schätzer in Ungleichung (5.3) *wird von ihnen nicht explizit gezeigt.*
Da sowohl die Varianz der Summe unabhängiger Zufallsvariablen in die Summe der Varianzen dieser Zufallsvariablen zerfällt, als auch der Varianzschätzer der Summe unabhängiger Zufallsvariablen in die Summe der Varianzschätzer dieser Zufallsvariablen (Lemma 5.2 und Satz 5.5), ist diese Aussage auch naheliegend und außerdem praktisch nur bedingt relevant, da für die zerlegte Situation der gewöhnliche Stichprobenvarianzschätzer verwendet werden kann. Die Bedeutung der Gleichheitsaussage liegt hauptsächlich darin, dass sie zeigt, dass die Normalisierungskonstante ohne genauere Kenntnis des Schätzers nicht verringert werden kann.

In wie weit der restliche Teil des Beweises nach Gleichung (5.4) *dem ähnelt, was sich Karlin und Rinott vorgestellt haben, ist mangels Angaben der Autoren nur schwer einschätzbar.*

Beweis von Satz 5.8. Unter den Voraussetzungen und mit den Notationen aus Satz 5.8 folgt:

$$\begin{aligned}
\frac{E(V_{\vec{k}}(S))}{K} &= \sum_{i_1\in I_1;\dots;i_m\in I_m} E(S_{\vec{k},\vec{i}} - \bar{S}_{\vec{k}})^2 = \sum_{i_1\in I_1;\dots;i_m\in I_m} E(S_{\vec{k},\vec{i}}(S_{\vec{k},\vec{i}} - \bar{S}_{\vec{k}}) - \bar{S}_{\vec{k}}(S_{\vec{k},\vec{i}} - \bar{S}_{\vec{k}})) \\
&= \left(\prod_{j=1}^{m} \binom{n_j}{p_j}\right)\left(E(S_{\vec{p}}^2) - E(S_{\vec{p}}\bar{S}_{\vec{k}})\right) \\
&= \left(\prod_{j=1}^{m} \binom{n_j}{p_j}\right)\left(E(S_{\vec{p}}^2) - E\left(S_{\vec{p}} \sum_{i_1\in I_1;\dots;i_m\in I_m} \frac{1}{\prod_{j=1}^{m}\binom{n_j}{k_j}} S_{\vec{k},\vec{i}}\right)\right) \\
&= \left(\prod_{j=1}^{m} \binom{n_j}{p_j}\right) E(S_{\vec{p}}^2) - \sum_{i_1\in I_1;\dots;i_m\in I_m} E(S_{\vec{p}}S_{\vec{k},\vec{i}}) \\
&= \left(\prod_{j=1}^{m} \binom{n_j}{p_j}\right) c_{\vec{p}} - \sum_{r_1=0}^{p_1}\cdots\sum_{r_m=0}^{p_m} c_{\vec{r}} \prod_{j=1}^{m} \binom{p_j}{r_j}\binom{n_j-p_j}{p_j-r_j}.
\end{aligned}$$

Die letzte Zeile folgt aus den Bemerkungen 17 und 15, Satz 5.5, Notation 6 und dem kombinatorischen Argument, dass es $\binom{p_j}{r_j}\binom{n_j-p_j}{p_j-r_j}$ Möglichkeiten gibt, im j-ten Variablenblock r_j Zufallsvektoren auszuwählen, die zu den ersten p_j in diesem Block gehören und die restlichen $p_j - r_j$ aus anderen Zufallsvektoren dieses Blocks.

Es bleibt zu zeigen, dass die Differenz zwischen Erwartungswert des Varianzschätzers und der Varianz der Zufallsvariable nichtnegativ ist. Das Symbol i für die Injektionen tritt im weiteren Verlauf nicht mehr auf und wird stattdessen an geeigneter Stelle als Multiindex wiederverwendet.

$$\begin{aligned}
&E(V_{\vec{k}}(S)) - \mathrm{Var}(S_{\vec{p}}) \\
=&K\left(\prod_{j=1}^{m} \binom{n_j}{p_j}\right) c_{\vec{p}} - K\left(\sum_{r_1=0}^{p_1}\cdots\sum_{r_m=0}^{p_m} c_{\vec{r}} \prod_{j=1}^{m} \binom{p_j}{r_j}\binom{n_j-p_j}{p_j-r_j}\right) - c_{\vec{p}} + c_{\vec{0}} \\
=&\left(K\left(\prod_{j=1}^{m} \binom{n_j}{p_j}\right) - 1\right)\left(\sum_{i_1=0}^{p_1}\cdots\sum_{i_m=0}^{p_m} E(A_{\vec{i}}^2) \prod_{j=1}^{m} \binom{p_j}{i_j}\right) \\
&- K\left(\sum_{r_1=0}^{p_1}\cdots\sum_{r_m=0}^{p_m}\sum_{i_1=0}^{p_1}\cdots\sum_{i_m=0}^{p_m} E(A_{\vec{i}}^2) \prod_{j=1}^{m} \binom{r_j}{i_j}\binom{p_j}{r_j}\binom{n_j-p_j}{p_j-r_j}\right) + E(A_{\vec{0}}^2) \\
=&E(A_{\vec{0}}^2) + \sum_{i_1=0}^{p_1}\cdots\sum_{i_m=0}^{p_m} E(A_{\vec{i}}^2)\Bigg(\\
&\left(K\left(\prod_{j=1}^{m} \binom{n_j}{p_j}\right) - 1\right)\left(\prod_{j=1}^{m} \binom{p_j}{i_j}\right) - K\left(\sum_{r_1=0}^{p_1}\cdots\sum_{r_m=0}^{p_m} \prod_{j=1}^{m} \binom{r_j}{i_j}\binom{p_j}{r_j}\binom{n_j-p_j}{p_j-r_j}\right)\Bigg)
\end{aligned} \tag{5.4}$$

Gleichung (5.4) ergibt sich durch Zusammenfassen der $c_{\vec{p}}$s und Ersetzung von $c_{\vec{p}}$ und $c_{\vec{r}}$ mittels Korollar 5.6.

Da alle $E(A_{\vec{i}}^2) \geq 0$, ist die Aussage gezeigt, wenn die Koeffizienten für alle $\vec{i}$ nichtnegativ sind.

Fallunterscheidung (1) $\vec{i} = \vec{0}$:

$$\begin{aligned}
&1 + \left(K\left(\prod_{j=1}^{m} \binom{n_j}{p_j} \right) - 1 \right) - K\left(\sum_{r_1=0}^{p_1} \cdots \sum_{r_m=0}^{p_m} \prod_{j=1}^{m} \binom{p_j}{r_j} \binom{n_j - p_j}{p_j - r_j} \right) \\
=&K\left(\left(\prod_{j=1}^{m} \binom{n_j}{p_j} \right) - \left(\sum_{r_1=0}^{p_1} \cdots \sum_{r_m=0}^{p_m} \prod_{j=1}^{m} \binom{p_j}{r_j} \binom{n_j - p_j}{p_j - r_j} \right) \right) = 0\,.
\end{aligned}$$

Das Ergebnis 0 folgt dabei über Induktion (die im nächsten Block nachgeholt wird) daraus, dass für jedes j die Anzahl der Möglichkeiten aus einer n_j elementigen Menge p_j Elemente auszuwählen, der Summe der Anzahlen der Möglichkeiten entspricht, aus den ersten p_j Elementen 0, bzw. $1, 2, \ldots, p_j$ Elemente auszuwählen und die jeweils restlichen Elemente aus den verbleibenden $n_j - p_j$ Elementen:

$$\text{für } m = 0 : \binom{n_j}{p_j} = \sum_{r_j=0}^{p_j} \binom{p_j}{r_j} \binom{n_j - p_j}{p_j - r_j}\,.$$

j ist 0 für $m = 0$, aber die obige Gleichung gilt generell und wird dementsprechend auch im nächsten Schritt benutzt werden. Im Induktionsschluss wird ein zusätzlicher Faktor f_j, der nicht von r abhängt, auf beiden Seiten eingebaut, der für den aktuellen Fall konstant 1 ist, aber in einem späteren Schritt des übergeordneten Beweis als $\binom{p_j}{i_j}$ benötigt wird. Im Induktionsanker ist er trivial und der Übersichtlichkeit halber ausgelassen worden.

$$\begin{aligned}
\text{IS: } \prod_{j=1}^{m+1} \binom{n_j}{p_j} f_j &= \left(\sum_{r_{m+1}=0}^{p_{m+1}} \binom{p_{m+1}}{r_{m+1}} \binom{n_{m+1} - p_{m+1}}{p_{m+1} - r_{m+1}} f_{m+1} \right) \left(\sum_{r_1=0}^{p_1} \cdots \sum_{r_m=0}^{p_m} \prod_{j=1}^{m} \binom{p_j}{r_j} \binom{n_j - p_j}{p_j - r_j} f_j \right) \\
&= \sum_{r_1=0}^{p_1} \cdots \sum_{r_m=0}^{p_m} \left(\sum_{r_{m+1}=0}^{p_{m+1}} \binom{p_{m+1}}{r_{m+1}} \binom{n_{m+1} - p_{m+1}}{p_{m+1} - r_{m+1}} f_{m+1} \right) \left(\prod_{j=1}^{m} \binom{p_j}{r_j} \binom{n_j - p_j}{p_j - r_j} f_j \right) \\
&= \sum_{r_1=0}^{p_1} \cdots \sum_{r_m=0}^{p_m} \left(\sum_{r_{m+1}=0}^{p_{m+1}} \binom{p_{m+1}}{r_{m+1}} \binom{n_{m+1} - p_{m+1}}{p_{m+1} - r_{m+1}} f_{m+1} \prod_{j=1}^{m} \binom{p_j}{r_j} \binom{n_j - p_j}{p_j - r_j} f_j \right) \\
&= \sum_{r_1=0}^{p_1} \cdots \sum_{r_{m+1}=0}^{p_{m+1}} \prod_{j=1}^{m+1} \binom{p_j}{r_j} \binom{n_j - p_j}{p_j - r_j} f_j
\end{aligned}$$

Damit ist der Fall $\vec{i} = \vec{0}$ vollständig bewiesen.

Fall (2) $\vec{i} \neq \vec{0}$:
Überspringe außerdem die Produktterme mit Binomialkoeffizient $= 0$.

Eine verwandte Betrachtung zur Partitionierung einer Menge in drei Teilmengen führt zu einer weiteren Anwendemöglichkeit der obigen Gleichheit und vereinfacht die anderen Koeffizienten: Die Anzahl der Möglichkeiten, aus einer p_j elementigen Menge eine r_j elementige und aus dieser wiederum eine i_j elementige auszuwählen, entspricht der Anzahl der Möglichkeiten aus den p_j Elementen i_j Elemente auszuwählen und aus den verbleibenden $p_j - i_j$ die restlichen $r_j - i_j$ Elemente:

$$
\begin{aligned}
&\left(K\left(\prod_{j=1}^{m}\binom{n_j}{p_j}\right)-1\right)\left(\prod_{j=1}^{m}\binom{p_j}{i_j}\right) - K\left(\sum_{r_1=i_1}^{p_1}\cdots\sum_{r_m=i_m}^{p_m}\prod_{j=1}^{m}\binom{r_j}{i_j}\binom{p_j}{r_j}\binom{n_j-p_j}{p_j-r_j}\right)\\
=&\left(K\left(\prod_{j=1}^{m}\binom{n_j}{p_j}\right)-1\right)\left(\prod_{j=1}^{m}\binom{p_j}{i_j}\right) - K\left(\sum_{r_1=i_1}^{p_1}\cdots\sum_{r_m=i_m}^{p_m}\prod_{j=1}^{m}\binom{p_j}{i_j}\binom{p_j-i_j}{r_j-i_j}\binom{n_j-p_j}{p_j-r_j}\right)\\
=&\left(K\left(\prod_{j=1}^{m}\binom{n_j}{p_j}\right)-1\right)\left(\prod_{j=1}^{m}\binom{p_j}{i_j}\right) - K\left(\sum_{r_1=0}^{p_1-i_1}\cdots\sum_{r_m=0}^{p_m-i_m}\prod_{j=1}^{m}\binom{p_j-i_j}{r_j}\binom{n_j-p_j}{p_j-i_j-r_j}\binom{p_j}{i_j}\right)\\
=&\left(K\left(\prod_{j=1}^{m}\binom{n_j}{p_j}\right)-1\right)\left(\prod_{j=1}^{m}\binom{p_j}{i_j}\right) - K\left(\sum_{r_1=0}^{p_1-i_1}\cdots\sum_{r_m=0}^{p_m-i_m}\prod_{j=1}^{m}\binom{(p_j-i_j)}{r_j}\binom{(n_j-i_j)-(p_j-i_j)}{(p_j-i_j)-r_j}\binom{p_j}{i_j}\right).
\end{aligned}
$$

Der Subtrahend besitzt dieselbe Form wie der Term im vorherigen Induktionsbeweis. Daher lässt er sich genauso umformen:

$$
\begin{aligned}
=&\left(K\left(\prod_{j=1}^{m}\binom{n_j}{p_j}\right)-1\right)\left(\prod_{j=1}^{m}\binom{p_j}{i_j}\right) - K\left(\prod_{j=1}^{m}\binom{p_j}{i_j}\binom{n_j-i_j}{p_j-i_j}\right)\\
=&K\left(\prod_{j=1}^{m}\binom{p_j}{i_j}\right)\left(\left(\prod_{j=1}^{m}\binom{n_j}{p_j}\right)-\left(\prod_{j=1}^{m}\binom{n_j-i_j}{p_j-i_j}\right)-K^{-1}\right).
\end{aligned}
$$

Das erste K und das erste Produkt können für die Abschätzung ausgelassen werden. Einsetzen von K in die restliche Gleichung führt anschließend zu:

$$
\begin{aligned}
&\left(\left(\prod_{j=1}^{m}\binom{n_j}{p_j}\right)-\left(\prod_{j=1}^{m}\binom{n_j-i_j}{p_j-i_j}\right)-\left(1-\max_{j\in\{1,\ldots,m\}}\frac{p_j}{n_j}\right)\prod_{j=1}^{m}\binom{n_j}{p_j}\right)\\
=&\left(\prod_{j=1}^{m}\binom{n_j}{p_j}\right)\left(1-\left(\frac{\prod_{j=1}^{m}\binom{n_j-i_j}{p_j-i_j}}{\prod_{j=1}^{m}\binom{n_j}{p_j}}\right)-\left(1-\max_{j\in\{1,\ldots,m\}}\frac{p_j}{n_j}\right)\right)\\
=&\left(\prod_{j=1}^{m}\binom{n_j}{p_j}\right)\left(\left(\max_{j\in\{1,\ldots,m\}}\frac{p_j}{n_j}\right)-\left(\prod_{j=1}^{m}\frac{\binom{n_j-i_j}{p_j-i_j}}{\binom{n_j}{p_j}}\right)\right).
\end{aligned}
$$

Das ausgeklammerte Produkt kann wieder vernachlässigt werden.

Betrachtungen zu den verbleibenden Termen: Es gibt für alle i_j mindestens so viele Möglichkeiten p_j Elemente aus einer n_j elementigen Menge zu wählen, wie es Möglichkeiten gibt $p_j - i_j$ Elemente aus einer $n_j - i_j$ elementigen Menge zu wählen, da sich der zweite Fall immer auf den ersten durch Hinzufügen neuer aber fest ausgewählter Elemente erweitert abschätzen lässt. Das Produkt der Quotienten der Binomialkoeffizienten wird daher mit jedem Faktor höchstens kleiner. Es reicht daher aus, die Nichtnegativität der Differenz für ein frei wählbares j nachzuweisen.

Außerdem reicht es für die Abschätzung aus, das Maximum der p_j/n_j durch den Quotienten für dieses j zu ersetzen. Offensichtlich kann das nicht für ein j mit $i_j = 0$ funktionieren. Da der Fall $i = 0$ schon bewiesen wurde, existiert nach Voraussetzung mindestens ein j mit $i_j > 0$. Es sei ein solches j gewählt.

Der Beweis wird hiermit auf den Fall eines einzigen Argumentenblocks zurückgeführt, d.h. einer Zufallsvariable der Form $S(X_1, \ldots, X_p)$. Der Index j wird im weiteren weggelassen, d.h.
$n = n_j, i = i_j, p = p_j$. Aus demselben kombinatorischen Argument, das nach dem letzten Gleichungsblock genannt wurde, ergibt sich, dass ein Beweis für $i = 1$ ausreicht:

$$
\frac{p}{n}-\frac{\binom{n-1}{p-1}}{\binom{n}{p}}=\frac{p}{n}-\frac{(n-1)!p!(n-p)!}{(p-1)!(n-p)!n!}=\frac{p}{n}-\frac{(n-1)!p!}{(p-1)!n!}=\frac{p}{n}-\frac{p}{n}=0\geq 0\,.
$$

Damit ist die Aussage, dass dieses Jackknife-Verfahren die Varianz in Erwartung nicht unterschätzt, bewiesen.

Es bleibt die Gleichheitsaussage für die Zufallsvariable T zu zeigen. Die ANOVA-Typ Zerlegung der Zufallsvariable T besteht aufgrund seiner Linearität und Unabhängigkeit der Zufallsvektoren gemäß Satz 5.1 und der zugehörigen Bemerkung 14 nur aus den A_i mit $< i, i > \leq 1$. Für $\vec{i} = 0$ wurde der Koeffizientenvergleich mit der Varianz bereits im Zuge des obigen Beweises durchgeführt. Es bleibt die Gleichheit für alle Koeffizienten mit $\vec{i}$ mit

$< i, i >$=1 zu zeigen. Hierfür bietet es sich an folgende Zeile aus obigen Beweis wieder aufzugreifen und gleich Null zu setzen:

$$\left(\prod_{j=1}^{m}\binom{n_j}{p_j}\right)\left(\left(\max_{j\in\{1,\dots,m\}}\frac{p_j}{n_j}\right)-\left(\prod_{j=1}^{m}\frac{\binom{n_j-i_j}{p_j-i_j}}{\binom{n_j}{p_j}}\right)\right)=0$$

$$\Leftrightarrow \frac{p_1}{n_1}-\prod_{j=1}^{m}\frac{\binom{n_j-i_j}{p_j-i_j}}{\binom{n_j}{p_j}}=0$$

$$\Leftrightarrow \frac{p_1}{n_1}-\frac{\binom{n_j-1}{p_j-1}}{\binom{n_j}{p_j}}=0 \qquad \text{für das } j \text{ mit } i_j=1\,.$$

Die letzte Gleichheit wurde bereits am Ende des Beweises zur Ungleichung gezeigt. Damit ist der gesamte Beweis abgeschlossen. □

Durch Anpassung der Zufallsvektoren kann T aus Satz 5.8 als allgemeiner linearer Schätzer verwendet werden.

Beispiel 6. Mit der Notation aus Satz 5.8 und für die Situation zweier unabhängiger Stichproben, $m = 2$, $Z_1, \dots, Z_{n_1}$ *i.i.d.* und $Y_1, \dots, Y_{n_2}$ *i.i.d.* schätzt T die Erwartungswertsdifferenz von Z_1 und Y_1, wenn $\forall_{i\in\{1,\dots,n_1\}} X_{i1} := Z_i/p_1$ und $\forall_{i\in\{1,\dots,n_2\}} X_{i2} := -Y_i/p_2$ gesetzt wird, wobei p_1 und p_2 beliebige Zahlen des Definitionsbereichs sein können.
Betrachtet sei der "leave one out" Fall $p_1 = 2$, $p_2 = 5$, $n_1 = 3, n_2 = 6$.
Hierfür gelten $\gamma = \frac{5}{6}$, $K^{-1} = \frac{1}{6}\binom{3}{2}\binom{6}{5} = 3$ und

$$S(X_{11}, X_{21}; X_{12}, X_{22}, X_{32}, X_{42}, X_{52}) := \frac{Z_1}{2}+\frac{Z_2}{2}+\frac{-Y_1}{5}+\frac{-Y_2}{5}+\frac{-Y_3}{5}+\frac{-Y_4}{5}+\frac{-Y_5}{5}\,.$$

I_1 lässt sich als $I_1 = \{i : \{1, 2\} \mapsto \{1, 2, 3\} | i(1) < i(2)\}$ wählen und I_2 als $I_2 = \{i : \{1, 2, 3, 4, 5\} \mapsto \{1, 2, 3, 4, 5, 6\} | i(1) < i(2) < i(3) < i(4) < i(5)\}$.
Damit ergibt sich:

$$\begin{aligned}\bar{S}_{1,1} &= \frac{1}{\binom{3}{1}\binom{6}{1}}\sum_{1\le\iota_1<\iota_2\le 3}\ \sum_{1\le\tau_1<\tau_2<\tau_3<\tau_4<\tau_5\le 6} S(X_{\iota_1 1}, X_{\iota_2 1}; X_{\tau_1 2}, X_{\tau_2 2}, X_{\tau_3 2}, X_{\tau_4 2}, X_{\tau_5 2})\\ &= \frac{1}{18}\left(\left(6\sum_{1\le\iota_1<\iota_2\le 3}\frac{Z_{\iota_1}+Z_{\iota_2}}{2}\right)-\left(3\sum_{1\le\iota_1<\iota_2<\iota_3<\iota_4<\iota_5\le 6}\frac{Y_{\iota_1}+Y_{\iota_2}+Y_{\iota_3}+Y_{\iota_4}+Y_{\iota_5}}{5}\right)\right)\\ &= \frac{1}{18}\left(\left(6\sum_{\iota=1}^{3}Z_\iota\right)-\left(3\sum_{\iota=1}^{6}Y_\iota\right)\right) = \bar{Z}-\bar{Y}\,.\end{aligned}$$

Die Umformung der Summen von der vorletzten auf die letzte Zeile lässt sich damit rechtfertigen, dass jeder Zufallsvektor genau 2 (bzw. 5) Mal in der jeweiligen Summe auftaucht, was wiederum genau dem Nenner entspricht. Für den Varianzschätzer zu $S_{2,5}$ gilt nun:

$$\begin{aligned}
V_{1,1}(S) &= \frac{1}{3} \sum_{1\le\iota_1<\iota_2\le 3} \sum_{1\le\tau_1<\cdots<\tau_5\le 6} \left(S(X_{\iota_1 1}, X_{\iota_2 1}; X_{\tau_1 2}, X_{\tau_2 2}, X_{\tau_3 2}, X_{\tau_4 2}, X_{\tau_5 2}) - \bar{S}_{1,1}\right)^2 \\
&= \frac{1}{3} \sum_{1\le\iota_1<\iota_2\le 3} \sum_{1\le\tau_1<\cdots<\tau_5\le 6} \left(\frac{Z_{\iota_1}+Z_{\iota_2}}{2} - \frac{Y_{\tau_1}+\cdots+Y_{\tau_5}}{5} - \bar{Z} + \bar{Y}\right)^2 \\
&= \frac{1}{3} \sum_{1\le\iota_1<\iota_2\le 3} \sum_{1\le\tau_1<\cdots<\tau_5\le 6} \Bigg(\left(\frac{Z_{\iota_1}+Z_{\iota_2}}{2} - \bar{Z}\right)^2 + \left(\frac{Y_{\tau_1}+\cdots+Y_{\tau_5}}{5} - \bar{Y}\right)^2 \\
&\qquad\qquad - 2\left(\frac{Z_{\iota_1}+Z_{\iota_2}}{2} - \bar{Z}\right)\left(\frac{Y_{\tau_1}+\cdots+Y_{\tau_5}}{5} - \bar{Y}\right)\Bigg).
\end{aligned}$$

Der dritte Summand wird 0, da kein ι in den Y-Term und kein τ in den Z-Term eingehen und die Terme einzeln aussummiert gleich 0 sind.
Ebenso lassen sich die beiden verbleibenden Summanden einzeln summieren:

$$\begin{aligned}
V_{1,1}(S) &= \sum_{1\le\iota_1<\iota_2\le 3} 2\left(\frac{Z_{\iota_1}+Z_{\iota_2}}{2} - \bar{Z}\right)^2 + \sum_{1\le\iota_1<\cdots<\iota_5\le 6} \left(\frac{Y_{\iota_1}+\cdots+Y_{\iota_5}}{5} - \bar{Y}\right)^2 \\
&= \sum_{\iota=1}^{3} 2\left(\frac{3\bar{Z}-Z_\iota}{2} - \bar{Z}\right)^2 + \sum_{\iota=1}^{6}\left(\frac{6\bar{Y}-Y_\iota}{5} - \bar{Y}\right)^2 \\
&= \sum_{\iota=1}^{3} \frac{1}{2}\left(\bar{Z}-Z_\iota\right)^2 + \frac{1}{5}\sum_{\iota=1}^{6}\frac{1}{5}\left(\bar{Y}-Y_\iota\right)^2 \\
E(V_{1,1}(S)) &= \operatorname{Var}(Z) + \frac{1}{5}\operatorname{Var}(Y) \\
&> \frac{1}{2}\operatorname{Var}(Z) + \frac{1}{5}\operatorname{Var}(Y) = \operatorname{Var}\left(\frac{Z_1+Z_2}{2}\right) + \operatorname{Var}\left(\frac{Y_1+\cdots+Y_5}{5}\right) = \operatorname{Var}(S_{2,5}).
\end{aligned}$$

Wäre $p_2 = 4$ gewählt worden, hätte sich eine genaue Varianzschätzung ergeben:
$\gamma = \frac{2}{3}$, $K^{-1} = \frac{1}{3}\binom{3}{2}\binom{6}{4} = 15$

$$S(X_{11}, X_{21}; X_{12}, X_{22}, X_{32}, X_{42}) = \frac{Z_1}{2} + \frac{Z_2}{2} + \frac{-Y_1}{4} + \frac{-Y_2}{4} + \frac{-Y_3}{4} + \frac{-Y_4}{4}$$

$$\begin{aligned}
\bar{S}_{1,2} &= \frac{1}{\binom{3}{1}\binom{6}{2}} \sum_{1\le\iota_1<\iota_2\le 3} \sum_{1\le\tau_1<\tau_2<\tau_3<\tau_4\le 6} S(X_{\iota_1 1}, X_{\iota_2 1}; X_{\tau_1 2}, X_{\tau_2 2}, X_{\tau_3 2}, X_{\tau_4 2}) \\
&= \frac{1}{45}\left(\left(15 \sum_{1\le\iota_1<\iota_2\le 3} \frac{Z_{\iota_1}+Z_{\iota_2}}{2}\right) - \left(3 \sum_{1\le\iota_1<\iota_2<\iota_3<\iota_4\le 6} \frac{Y_{\iota_1}+Y_{\iota_2}+Y_{\iota_3}+Y_{\iota_4}}{4}\right)\right) \\
&= \frac{1}{45}\left(\left(15\sum_{\iota=1}^{3} Z_\iota\right) - \left(\frac{15}{2}\sum_{\iota=1}^{6} Y_\iota\right)\right) = \bar{Z} - \bar{Y}\,.
\end{aligned}$$

Bei der Summenumformung zur letzten Zeile wird wieder beachtet, dass alle Zufallsvektoren Y_ι gleichhäufig auftauchen. Y_1 fehlt in dem Summand genau dann, wenn $\iota_1 > 1$. Die Häufigkeit hierfür entspricht der Anzahl der Möglichkeiten aus der Menge $\{2,3,4,5,6\}$ vier verschiedene Zahlen auswählen zu können. Jeder Summand tritt demnach mit der Häufigkeit $\binom{6}{4} - \binom{5}{4} = 10$ auf. Für die Varianzschätzung ergibt sich:

$$\begin{aligned}
V_{1,2}(S) &= \frac{1}{15} \sum_{1\le\iota_1<\iota_2\le 3} \sum_{1\le\tau_1<\cdots<\tau_4\le 6} \left(S(X_{\iota_1 1}, X_{\iota_2 1}; X_{\tau_1 2}, X_{\tau_2 2}, X_{\tau_3 2}, X_{\tau_4 2}) - \bar{S}_{1,1}\right)^2 \\
&= \frac{1}{15} \sum_{1\le\iota_1<\iota_2\le 3} \sum_{1\le\tau_1<\cdots<\tau_4\le 6} \left(\frac{Z_{\iota_1}+Z_{\iota_2}}{2} - \frac{Y_{\tau_1}+\cdots+Y_{\tau_4}}{4} - \bar{Z} + \bar{Y}\right)^2 \\
&= \frac{1}{15} \sum_{1\le\iota_1<\iota_2\le 3} \sum_{1\le\tau_1<\cdots<\tau_4\le 6} \left(\left(\frac{Z_{\iota_1}+Z_{\iota_2}}{2} - \bar{Z}\right)^2 + \left(\frac{Y_{\tau_1}+\cdots+Y_{\tau_4}}{4} - \bar{Y}\right)^2\right) \\
&= \sum_{1\le\iota_1<\iota_2\le 3} \left(\frac{Z_{\iota_1}+Z_{\iota_2}}{2} - \bar{Z}\right)^2 + \frac{3}{15} \sum_{1\le\iota_1<\cdots<\iota_4\le 6} \left(\frac{Y_{\iota_1}+\cdots+Y_{\iota_4}}{4} - \bar{Y}\right)^2 \\
&= \frac{1}{2}\sum_{\iota=1}^{3} \frac{1}{2}\left(\bar{Z} - Z_\iota\right)^2 + \frac{3}{15}\sum_{\iota=1}^{6}\sum_{\tau=\iota+1}^{6} \left(\frac{6\bar{Y} - Y_\iota - Y_\tau}{4} - \bar{Y}\right)^2 .
\end{aligned}$$

Der zweite Term lässt sich durch Verdopplung des Wertebereichs von τ (ausgeglichen durch Korrekturfaktor $\frac{1}{2}$) und anschließender 0-Addition des "$\iota = \tau$"-Falls vereinfachen:

$$\begin{aligned}
\frac{3}{30}\sum_{\iota=1}^{6}\sum_{\tau\neq\iota}\left(\frac{\bar{Y}-Y_\iota}{4}+\frac{\bar{Y}-Y_\tau}{4}\right)^2 &= \frac{1}{160}\left(\sum_{\iota=1}^{6}\sum_{\tau=1}^{6}\left((\bar{Y}-Y_\iota)+(\bar{Y}-Y_\tau)\right)^2 - \sum_{\iota=1}^{6}\left(2\bar{Y}-2Y_\iota\right)^2\right) \\
&= \frac{12}{160}\sum_{\iota=1}^{6}(\bar{Y}-Y_\iota)^2 - \frac{1}{160}\sum_{\iota=1}^{6}\left(2\bar{Y}-2Y_\iota\right)^2 = \frac{1}{20}\sum_{\iota=1}^{6}(\bar{Y}-Y_\iota)^2 .
\end{aligned}$$

Und damit gilt: $$V_{1,2}(S) \;=\; \frac{1}{2}\sum_{\iota=1}^{3}\frac{1}{2}\left(\bar{Z}-Z_\iota\right)^2 \;+\; \frac{1}{4}\sum_{\iota=1}^{6}\frac{1}{5}(\bar{Y}-Y_\iota)^2\,,$$

$$E(V_{1,2}(S)) = \frac{1}{2}\text{Var}(Z) + \frac{1}{4}\text{Var}(Y) = \text{Var}\left(\frac{Z_1 + Z_2}{2}\right) + \text{Var}\left(\frac{Y_1 + \ldots + Y_4}{4}\right) = \text{Var}(S_{2,4})\,.$$

5.2.1. Genauigkeit der Varianzschätzung

Beispiel 6 zeigt eine aufwendige Berechnung der erwarteten Varianzschätzung und deren Abweichung von der tatsächlichen Varianz. Das Beispiel liegt in einer Klasse von Schätzern, für die sich theoretische Aussagen zur Genauigkeit der Varianzschätzung formulieren lassen:

Satz 5.9 (Bias des Varianzschätzers $V_{\vec{k}}(T)$ bei ungleichen p_j/n_j Werten):
Es seien $\forall_{j\in\{1,\ldots,m\}} \widetilde{T}_j := X_{1j} + \ldots + X_{p_j j}$ m *unabhängige Zufallsvariablen.*
Der Bias des Schätzers $V_{\vec{k}}(T)$ *für die Varianz von* $T = \sum_{j=1}^{m} \widetilde{T}_j$ *aus Satz 5.8 ist dann:*

$$EV_{\vec{k}}(T) - VarT = \sum_{j=1}^{m} \frac{\left(k_j/n_j\right) - \min_{l\in\{1,\ldots,m\}}\left(k_l/n_l\right)}{\min_{l\in\{1,\ldots,m\}}\left(k_l/n_l\right)} Var\widetilde{T}_j\,.$$

Der Faktor, um den die Varianz des Schätzers T *in Erwartung gestreckt wird, ist:*

$$\frac{EV_{\vec{k}}(T)}{VarT} = \sum_{j=1}^{m} \frac{\left(k_j/n_j\right)}{\min_{l\in\{1,\ldots,m\}}\left(k_l/n_l\right)} \frac{Var\widetilde{T}_j}{VarT}\,.$$

Bemerkung 21:
Die $\widetilde{T}_j$ *aus Satz 5.9 besitzen alle die Struktur von* T *aus Satz 5.7, bzw. der einzelnen Blöcke aus Satz 5.8.*

Beweis von Satz 5.9. Der Beginn des Beweises ist etwas allgemeiner gehalten und wird für Schätzer durchgeführt, die in Summen von Funktionen von den in Satz 5.8 definierten Blöcken von Zufallsvektoren zerlegt werden können. Diese Blöcke seinen hier, wie in Satz 5.9 gefordert, untereinander unabhängig.

Für die zugehörigen Zufallsvariablen $\widetilde{S}$ gilt: $S_{\vec{p}} =: \sum_{j=1}^{m} \widetilde{S}_j$.

Für die $\bar{S}$ Notation aus Satz 5.8 folgt unmittelbar: $S_{\vec{p}} = \sum_{j=1}^{m} \widetilde{S}_j \Rightarrow \bar{S}_{\vec{k}} = \sum_{j=1}^{m} \bar{\widetilde{S}}_j$.

Für beliebige $\vec{\tau} \in I_1 \times \ldots \times I_m$ gilt:

$$E\bar{S}_{\vec{k}} = \frac{1}{\prod_{j=1}^{m}\binom{n_j}{k_j}} \sum_{i_1\in I_1,\ldots,i_m\in I_m} ES_{\vec{k},\vec{i}} = ES_{\vec{k},\vec{\tau}} = ES_{\vec{p}}\,.$$

Für Differenzen dieser Terme entspricht, wegen der Zentriertheit, das zweites Moment der Varianz. Der Erwartungswert des Schätzers V lässt sich nun vereinfachen:

$$
\begin{aligned}
EV_{\vec{k}}(S) =& \frac{1}{(1-\gamma)\prod_{j=1}^{m}\binom{n_j}{k_j}} \sum_{i_1\in I_1,\ldots,i_m\in I_m} E\big(S_{\vec{k},\vec{i}} - \bar{S}_{\vec{k}}\big)^2 \\
=& \frac{1}{(1-\gamma)\prod_{j=1}^{m}\binom{n_j}{k_j}} \sum_{i_1\in I_1,\ldots,i_m\in I_m} \operatorname{Var}\big(S_{\vec{k},\vec{i}} - \bar{S}_{\vec{k}}\big) = \frac{1}{(1-\gamma)} \operatorname{Var}\big(S_{\vec{p}} - \bar{S}_{\vec{k}}\big) \\
=& \frac{1}{(1-\gamma)} \sum_{j=1}^{m} \operatorname{Var}\Big(\widetilde{S}_j - \bar{\widetilde{S}}_j\Big) = \frac{1}{(1-\gamma)} \sum_{j=1}^{m} \left(1 - \frac{p_j}{n_j}\right) EV_{k_j}\big(\widetilde{S}_j\big) \qquad (5.5) \\
& \text{Für } T = S \text{ gilt nun:} \\
=& \frac{1}{(1-\gamma)} \sum_{j=1}^{m} \left(1 - \frac{p_j}{n_j}\right) \operatorname{Var}\big(\widetilde{T}_j\big) = \sum_{j=1}^{m} \frac{1 - \frac{p_j}{n_j}}{1 - \max_{l\in\{1,\ldots,m\}}\left(\frac{p_l}{n_l}\right)} \operatorname{Var}\big(\widetilde{T}_j\big) \\
=& \sum_{j=1}^{m} \frac{\big(k_j/n_j\big)}{\min_{l\in\{1,\ldots,m\}}\big(k_l/n_l\big)} \operatorname{Var}\big(\widetilde{T}_j\big)\,.
\end{aligned}
$$

Gleichung 5.5 gilt, da der Quer-Opperator ($\bar{}$), aus Satz 5.8 linear ist. Die Blöcke können daher separiert werden. Weil sie unabhängig sind, lässt sich die Summe anschließend vollständig aus der Varianz herausziehen.

Die Formel zur Varianzstreckung ist damit offensichtlich. Die Formel für den Bias folgt aus $\operatorname{Var} T = \sum_{j=1}^{m} \operatorname{Var} \widetilde{T}_j$. □

Simulationsansatz: Für jeden praktischen Anwendungsfall mit bekannt angenommener Verteilung lässt sich eine Varianzüberschätzung mit Simulationen quantifizieren. Ein approximativer Algorithmus ist im Folgenden angegeben und wird auf Beispiel 6 angewandt.

Simulationsroutine

```
getrandpermcols = function(n,p,reps=1000)
  replicate(reps, unlist(Map(function(a,b)a+b,
            Map(sample.int,n,p), head(cumsum(c(0,n)),-1))))

V = function(k,n,x,S,perms, Sbar=mean(apply(perms,2,function(idx)S(x[idx]))))
      mean((apply(perms,2,function(idx)S(x[idx]))-Sbar)**2) / min(k/n)

Smeandiff.2.5 = function(x) mean(x[1:2])-mean(x[3:7])

set.seed(123)
Vwert = mean(replicate(1000, V(c(1,1), c(3,6), rnorm(9),
                    Smeandiff.2.5, getrandpermcols(c(3,6),c(2,5),1000))))
Varianz = var(replicate(1000, Smeandiff.2.5(rnorm(7))))
Vwert / Varianz
```

Simulations-Ergebnisse:

$$\begin{aligned}\text{Vwert} &= 1.2075 \approx 1.2 = 1\,\text{Var}(N_{(0,1)}) + \tfrac{1}{5}\,\text{Var}(N_{(0,1)}) \\ \text{Varianz} &= 0.7023861 \approx 0.7 = \text{Var(Smeandiff.2.5)} \\ \text{Vwert/Varianz} &= 1.71914\end{aligned}$$

Simulationsumgebung: Wenn für einen Versuch ein Abstandsmaß (oder eins von mehreren in Frage kommenden) verwendet werden soll, wäre dieses R-Skript nur bedingt geeignet um eine geeignete Kombination eines Varianzschätzers und einer Statistik zu finden. Eine vollständige Simulationsumgebung für Äquivalenzteste, die von mir im Rahmen dieser Dissertation erstellt wurde, befindet sich im Supplement (DOI 10.13140/RG.2.1.3287.6407) von Antweiler *et al.* (2017). Die Simulationsumgebung wurde entwickelt, um beliebige Kombinationen aus Varianzschätzer und Abstandsmaß und darauf beruhender Statistik in umfangreichen Szenarien mit kleinen Stichprobengrößen zu untersuchen. Design-Ziele waren Korrektheit, Reproduzierbarkeit, Dokumentation und Geschwindigkeit. Die Simulationsumgebung besteht aus verschiedenen Bausteinen. Die Hauptroutinen sind in C geschrieben. Einige Abstände, Abstandswertkombinierer, Transformationen und Varianzschätzer (einschließlich dem oben definierten mit vollständigen nicht-zufälligen Permutationen) sind bereits enthalten. Weitere lassen sich in Form von kompilierten C-Funktionen hinzulinken, ohne dass das Programm selbst recompiliert werden muss. Die Parameter für die Simulationen werden in einer `SQLite3` Datenbank definiert. Beim Ausführen des Programms werden diese mit den Simulationsergebnissen, einem Time-Stamp, der Versionsnummer und Hash-Werten des Simulationsprogramms und von eventuell verwendeter Daten verknüpft und in dieser Datenbank hinterlegt. Um die Simulationen durch die Parameter flexibel steuern zu können, wurde eine Interpreter-Routine für elementare Matrix- und Vektoralgebra als Bestandteil des C-Programms implementiert. Reale Daten können ebenfalls in einer `SQLite3`-Datenbank hinterlegt und für die Simulationen oder Daten-Analysen benutzt werden. Zur Integration der Daten in die Datenbank wurde ein `awk`-Skript geschrieben, dass Daten aus `csv`-Dateien auslesen kann. Front-Ends für `SQLite3`-Datenbanken können als Benutzeroberflächen für die Definition der Simulations-Szenarien verwendet werden. Installation und Benutzung des Programms ist in der Datei "README" beschrieben, die dem Programm beiliegt.

Bemerkung 22:
Aus einer Konservativität eines Varianzschätzers folgt noch nicht die Konservativität des darauf aufbauenden Tests. Für den Datensatz, für den dieses Jackknifeverfahren gewählt wurde, stellte sie jedoch die entscheidende Komponente dar. Bei einem Test, der eine breite Wahl an Teststatistiken erlauben soll, empfiehlt es sich zusätzlich, den Test für die gewählte Statistik in Simulationsstudien mit theoretischen und resampleten realen Daten zu überprüfen.

6. Der Testablauf insgesamt

Zusammenfassend lässt sich der vollständige Testablauf folgend darstellen:

Messungen: (Abschnitt 4.1)
Für den Test werden drei verschiedene Gruppen beprobt:

- Eine Behandlungsgruppe, über die eigentlich die Aussage getroffen werden soll.
- Eine Kontrollgruppe, die gegen die Behandlungsgruppe verglichen wird. Aus der Kontrollgruppe kann auch eine zusätzliche Stichprobe gezogen werden, die für die Schätzung der Äquivalenzschranke verwendet wird. Es kann aber auch dieselbe Probe für die Schrankenbestimmung wie für den Vergleich zur Behandlungsgruppe verwendet werden.
- Eine dritte Gruppe, die nur zur Schätzung der Äquivalenzschranke benötigt wird. Sie wurde einer anderen Behandlung unterzogen, deren Effekte als akzeptabel angesehen werden. Falls bereits ein Schrankenwert bekannt ist, kann auf diese Gruppe verzichtet werden.

Beispiel 7. In einer Kontrollgruppe "x" liegen die beiden 2-dimensionalen Messungen $x_1 = (3, 5)$ und $x_2 = (2, 4)$ vor. In einer Behandlungsgruppe "y" die drei Messungen $y_1 = (4, 5)$ $y_2 = (4, 7)$ und $y_3 = (2, 3)$. In der dritten Gruppe "z" die zwei Messungen $z_1 = (6, 8)$ und $z_2 = (5, 9)$.

Paarweise Abstandsmaße: (Abschnitt 4.2.1 und Anhang A)
Um die Information der hochdimensionalen Daten geeignet zusammenzufassen, wird ein sinnvolles Abstandsmaß benutzt, für das die Teststatistik später annähernd normalverteilt erscheint. Für die Mikrobiom-Daten empfiehlt sich der relative Bray-Curtis Abstand (Definition 4.1, Seite 25), der proportional zum Abstand der 1-Norm ist. Mit dem gewählten Abstand werden alle paarweisen Abstände der Behandlungs und Kontrollgruppe berechnet, so dass eine Matrix aus Abstandswerten entsteht. Je nach verwendeter Teststatistik werden auch die Abstände innerhalb jeder der Gruppen berechnet. Ebenso wird mit der dritten Gruppe und der Kontrollgruppe verfahren.

Fortsetzung von Beispiel 7:
Als Abstandsmaß werde der quadratische Maximumsabstand $d(a,b) \mapsto \max((a_1-b_1)^2,(a_2-b_2)^2)$ verwendet. Damit berechnen sich die Abstände:

$$\begin{pmatrix} d(x_1,y_1) & d(x_1,y_2) & d(x_1,y_3) \\ d(x_2,y_1) & d(x_2,y_2) & d(x_2,y_3) \end{pmatrix}$$
$$= \begin{pmatrix} \max\left((3-4)^2,(5-5)^2\right) & \max\left((3-4)^2,(5-7)^2\right) & \max\left((3-2)^2,(5-3)^2\right) \\ \max\left((2-4)^2,(4-5)^2\right) & \max\left((2-4)^2,(4-7)^2\right) & \max\left((2-2)^2,(4-3)^2\right) \end{pmatrix} = \begin{pmatrix} 1 & 4 & 4 \\ 4 & 9 & 1 \end{pmatrix}$$

und

$$\begin{pmatrix} d(x_1,z_1) & d(x_1,z_2) \\ d(x_2,z_1) & d(x_2,z_2) \end{pmatrix} = \begin{pmatrix} \max\left((3-6)^2,(5-8)^2\right) & \max\left((3-5)^2,(5-9)^2\right) \\ \max\left((2-6)^2,(4-8)^2\right) & \max\left((2-5)^2,(4-9)^2\right) \end{pmatrix} = \begin{pmatrix} 9 & 16 \\ 16 & 25 \end{pmatrix}$$

Teststatistik: (Abschnitt 4.2.2)
Die Abstandsmatrix aus Behandlungs- und Kontroll-Proben wird zu einer einzelnen Zahl zusammengefasst. Eine einfache und gute Möglichkeit hierfür ist Gleichung (4.1) von Seite 27. Diese kann direkt als Teststatistik verwendet oder vorher noch transformiert werden, um die Verteilung besser an eine Normalverteilung anzupassen. Die Transformation sollte einfach und intuitiv sein, wenn die Stichprobengröße klein ist. Bei quadratischen Abständen bietet sich eine Wurzeltransformation an, die gegebenenfalls an den möglichen Wertebereich angepasst werden muss. Die Äquivalenzschranke ergibt sich durch dieselbe Rechnung mit der Abstandsmatrix aus der dritten Gruppe und der Kontrollgruppe.

Fortsetzung von Beispiel 7:
Für die Teststatistik wird Gleichung (4.1) verwendet

$$r_b = \frac{1}{n_x n_y} \sum_{i=1}^{n_x} \sum_{j=1}^{n_y} d(x_i,y_j) = \frac{1}{2\cdot 3}(1+4+4+4+9+1) = \frac{23}{6}$$

und Wurzel-transformiert $T(x,y) = \sqrt{r_b} = \sqrt{23/6} \approx 1.96$.
Die Äquivalenzschranke berechnet sich zu

$$\sqrt{T(x,z)} = \sqrt{\tfrac{1}{2\cdot 2}(9+16+16+25)} = \sqrt{66/4} \approx 4.06\,.$$

Varianzschätzung: (Abschnitt 5.2)
Viele Verfahren zur Schätzung der Varianz der Teststatistik sind bei kleinen Stichproben sehr liberal. Eine Ausnahme stellt das Jackknife-Prinizip dar. Eine Version für mehrere Gruppen wurde hier an nicht-balancierte Designs angepasst. Ein Jackknife-Verfahren schätzt die Varianz einer Statistik, in der nicht alle Stichprobenelemente eingesetzt sind. Oft kann diese Varianz als konservative Schätzung der vollständigen Teststatistik interpretiert werden. In einigen Fällen lassen sich auch Korrekturfaktoren für die Übertragung auf die vollständige Teststatistik angeben - siehe Bemerkung 19.

Fortsetzung von Beispiel 7:
In dem 2-Gruppen Jackknife-Verfahren sollen aus der Behandlungsgruppe jeweils 2 und aus der Kontrollgruppe jeweils eine Probe verwendet werden. Im Sinne von Satz 5.8 ist somit $n = (2, 3)$, $p = (1, 2)$, $k = (1, 1)$, $\gamma = 2/3$ und $K = 3\binom{2}{1}\binom{3}{2} = 18$, wenn "x" den ersten Block und "y" den zweiten bezeichnet. Wie das Verfahren ausgeführt wird wurde bereits ausführlich in Beispiel 6 demonstriert. Daher werden hier nur die relevanten Zwischen- und Endergebnisse aufgelistet: Es ergibt sich

$$\begin{aligned}\bar{s}_{1,1} =& \frac{1}{\binom{2}{1}\binom{3}{2}}\Big(\sqrt{\tfrac{1}{2}(9+1)} + \sqrt{\tfrac{1}{2}(4+1)} + \sqrt{\tfrac{1}{2}(4+9)} + \\ & \sqrt{\tfrac{1}{2}(4+4)} + \sqrt{\tfrac{1}{2}(1+4)} + \sqrt{\tfrac{1}{2}(1+4)}\Big) \approx 1.92\end{aligned}$$

und schließlich als konservative Varianzschätzung für $T(X, Y)$:

$$\begin{aligned}v_{1,1} = 3\Big(&\big(\sqrt{5} - 1.92\big)^2 + \big(\sqrt{5/2} - 1.92\big)^2 + \big(\sqrt{13/2} - 1.92\big)^2 + \\ &\big(2 - 1.92\big)^2 + \big(\sqrt{5/2} - 1.92\big)^2 + \big(\sqrt{5/2} - 1.92\big)^2\Big) \approx 3 \cdot 0.85 \approx 5.76\,.\end{aligned}$$

Konfidenzintervall: Mit dem Wert der Teststatistik und ihrer Varianz lässt sich ein normalverteiltes Konfidenzintervall angeben. Um Abweichungen von der Normalverteilung abzupuffern oder an dieser Stelle keine Liberalität zu erzeugen, kann die Formel für t-Verteilungen verwendet werden. Aus diesem Gesichtspunkt kann der Freiheitsgrad auch großzügig gewählt werden, beispielsweise als Anzahl der Stichproben beider Gruppen.

Fortsetzung von Beispiel 7:
Mit der Formel eines 1-seitigen NV-Konfidenzintervalls $(\infty, \hat{T} + 1.64\sqrt{\widehat{\mathrm{Var}(T)}})$ zum Niveau $\alpha = 0.05$ ergibt sich die Abschätzung $(\infty, 1.96 + 1.64\sqrt{5.76}) = (\infty, 5.90)$.

Testentscheid: Wenn der obere Rand dieses Konfidenzintervalls die Äquivalenzgrenze nicht überschreitet, ist der Test signifikant.

Fortsetzung von Beispiel 7:
In dem Beispiel liegt die Äquivalenzgrenze 4.06 innerhalb des Konfidenzintervalls $(\infty, 5.90)$. Der Test wäre daher nicht signifikant. Die Stichprobe war allerdings auch sehr klein.

6.1. Version mit Schranke pro Stichprobenelement

Die Verwendung einer ökologisch bedeutsamen Schranke passt hauptsächlich zu der Idee, dass diese Schranke höchstens selten überschritten werden sollte, und die Fälle mit Überschreitungen schlimm sein könnten. Das vorgestellte Verfahren überprüft hingegen, ob (über Stichproben) gemittelte Werte unterhalb der Schranke liegen. Bei einem kleinem Standardfehler kann ein Großteil der Stichprobe jedoch oberhalb der Schranke liegen. Die Stichprobengrößen sind momentan klein genug, dass das unwahrscheinlich ist, es lässt sich aber auch ein Verfahren verwenden, dass genau für das Problem der individuellen Einhaltung der Schranken konstruiert ist.

Hierfür werden die Einträge der Abstandsmatrix mit Hilfe der Schranke in dichotome Werte transformiert: 1 für überschritten und 0 für eingehalten. Der Test kann dann mit dieser neuen Abstandsmatrix anstatt der ursprünglichen durchgeführt werden. Es wird dann allerdings eine weitere Äquivalenz-Schranke benötigt - diesmal für den Anteil der Stichprobenpaare, denen man zugestehen will, die erste Schranke zu überschreiten.

Einfacher wäre es, fast nur Proben aus der Behandlungsgruppe zu sammeln und für diese anschließend Abstandswerte zu einer einzigen Kontrollprobe zu berechnen. Die dichotomen Abstandswerte könnten dann als unabhängige Messwerte aus einer Bernoulli-Verteilung betrachtet und mit einem Binomialtest überprüft werden. Die zweite Schranke (der Wahrscheinlichkeitsparameter des Tests) wäre dabei etwas intuitiver und die Test-Durchführung weniger aufwendig und verteilungsunabhängig (im Sinne der ursprünglichen Daten) durchführbar. Alternativ zur Verwendung einer einzigen Kontrollprobe können auch jeweils die Abstände zu allen Kontrollproben berechnet und für jede Probe der Behandlungsgruppe deren Maximum genommen werden. Letzteres wird in folgendem Beispiel vorgeführt.

Beispiel 8. Es werden die Zahlen aus Beispiel 7 verwendet. Als Schranke wird das Minimum der Abstandswerte zwischen Kontroll- und dritter Gruppe verwendet: 9 . In der Behandlungsgruppe sind die größten Abstände zu einem Wert der Kontrollgruppe jeweils:

4, 9 und 4. Ein Wert auf der Schranke soll, in diesem Beispiel, bereits als Überschreitung angesehen werden. Als dichotome Abstandswerte ergeben sich: 0, 1 und 0. In 2 der 3 Fälle liegt der Wert unterhalb der Schranke. Die Teststatistik für den Binomialtest ist beispielsweise 1/3 für den Anteil der Überschreitungen. Wenn 10% Überschreitungen akzeptiert werden würden, wäre der p-Wert des 1-seitigen Tests:

$$B_{3,0.1}(\{0,1\}) = \binom{3}{0} 0.1^0 \cdot 0.9^3 + \binom{3}{1} 0.1^1 \cdot 0.9^2 = 1 \cdot 1 \cdot 0.9^3 + 3 \cdot 0.1 \cdot 0.9^2 = 0.972\,.$$

7. Simulationen

Grundlage der Simulationen ist das Abstandsmaß $\forall_{a,b\in\mathbf{R}^p}\, d(a,b) := \max\limits_{1\leq k\leq p} (a_k - b_k)^2$.

Als Transformation einer Teststatistik wird die Funktion $\forall_{r\in\mathbf{R}} f(r) := \text{sign}(r)\sqrt{|r|}$

verwendet (siehe Abschnitt A). Dabei handelt es sich um eine streng monotone Fortsetzung der gewöhnlichen Wurzelfunktion auf den negativen Bereich. Diese ist notwendig, wenn die Teststatistik negativ werden kann und auch die Streuungsinformation von negativen Zahlen für die Analyse genutzt werden soll. Die präsentierten Ergebnisse beruhen allerdings auf der Statistik $S = r_b$ aus Gleichung (4.1), die bei dem verwendeten Abstandsmaß nicht negativ werden kann. Für diese Statistik und die beiden Stichproben $x_1, \ldots, x_{n_x} \in \mathbf{R}^p$ und $y_1, \ldots, y_{n_y} \in \mathbf{R}^p$, $p \in \mathbf{N}_{>0}$ ergibt sich somit insgesamt:

$$T(x,y) := f(S(x,y)) = \sqrt{\frac{1}{n_x n_y} \sum_{i=1}^{n_x} \sum_{j=1}^{n_y} \max_{1\leq k\leq p} \left((x_{ik} - y_{jk})^2\right)} \quad .$$

7.1. Überdeckungsraten

Ein Konfidenzintervall kann zur Konstruktion eines Äquivalenztests genutzt werden (Abschnitt 3.2). Wenn das verwendete Konfidenzintervall sein Niveau einhält, hält auch der Äquivalenztest sein Niveau ein. Ein Äquivalenztest kann sein Niveau auch einhalten, ohne dass sein Konfidenzintervall es einhält. Das wird an dieser Stelle ignoriert. Stattdessen wird in diesem Kapitel in Simulationen kontrolliert, ob Konfidenzintervalle ihre Überdeckungswahrscheinlichkeiten für verschiedene Werte einhalten. Die Werte sind jeweils die Erwartungswerte der Teststatistik T . Allerdings wurde die Konservativität bei Betrachtungen auch an parametrischen Bootstrap-Stichproben (multinomialen Zufallsstichproben, mit den realen Messdaten als Wahrscheinlichkeitsparametern) direkt anhand des verwendeten Äquivalenztests überprüft.

Definition 7.1 (Überdeckungsrate). Die relative Häufigkeit der Realisationen eines Konfidenzintervalls, die den zu schätzenden Wert enthalten, heißt Überdeckungsrate. Sie schätzt die Überdeckungswahrscheinlichkeit des Konfidenzintervalls.

Jede Überdeckungswahrscheinlichkeit wird in diesem Kapitel mit 10000 Simulationen pro Parameterkombination geschätzt. Da die zu schätzenden Werte zwar eindeutig von dem Simulationsszenario abhängen, aber nicht bekannt sind, müssen auch diese geschätzt werden. Hierfür wird der Mittelwert der Schätzer T über alle 10000 Simulationen verwendet. Für jedes der 10000 Realisationen eines Konfidenzintervalls $C(x, y)$ wird dann geprüft, ob dieser Mittelwert enthalten ist. x und y sind die simulierten Werte-Matrizen eines Simulationsschritts der beiden Gruppen.

Überdeckungsrate = $\frac{\text{Anzahl } \bar{T} \in C(x,y)}{10000}$.

Die Konfidenzintervalle sind auf ein Niveau von 0.95 angesetzt.

7.1.1. Univariate Simulationen

Die erste Simulationsreihe wurde für eindimensionale Endpunkte angefertigt. Dabei wurden 2 unabhängige Gruppen verwendet. Beide Gruppen waren stets gleich groß: $n := n_x = n_y$. Es wurden alle Simulationen sowohl mit Gruppengröße $n = 5$ als auch mit Gruppengröße $n = 10$ durchgeführt. Die Mittelwerte der beiden Gruppen wurden gegeneinander verschoben. Beide Gruppen sind normalverteilt und haben dieselbe Varianz.

Das Abstandsmaß und die Tranformation vereinfachen sich im univariaten Fall für die verwendete Statistik zu

$$d(a, b) := (a - b)^2$$
$$f = \sqrt{\ } .$$

Im Fall $\mu_x \gg \mu_y$ gilt für den Schätzer:

$$\left|\bar{X} - \bar{Y}\right| \approx \bar{X} - \bar{Y} \quad \sim \text{normalverteilt} . \tag{7.1}$$

Die Relation $\gg$ ist hierbei im Verhältnis zur Varianz zu sehen.

Bemerkung 23:
Nichtparametrische Intervallschätzer haben sich als ungeeignet herausgestellt.

- *Alle getesteten Bootstrapvarianten (auch parametrische) waren antikonservativ mit Überdeckungsraten bei ungünstigen Parametern* $< 80\%$.
- *Hodges-Lehmann Konfidenzintervalle lagen unter* $< 30\%$.

Ergebnisse der Simulationsstudie für das 2-Gruppen-Jackknifeverfahren haben gezeigt, dass die Überdeckungsrate bei Verschiebungen um weniger als eine Standardabweichung über 95% liegt. Bei Verschiebungen um weniger als zwei Standardabweichungen liegt sie über 93.5% für $n = 5$ und über 94% für $n = 10$. Bei größeren Verschiebungen liegt sie ungefähr bei 95% (Abb. 7.1). Letzteres entspricht den aufgrund von Gleichung (7.1) zu erwartenden theoretischen Ergebnissen.

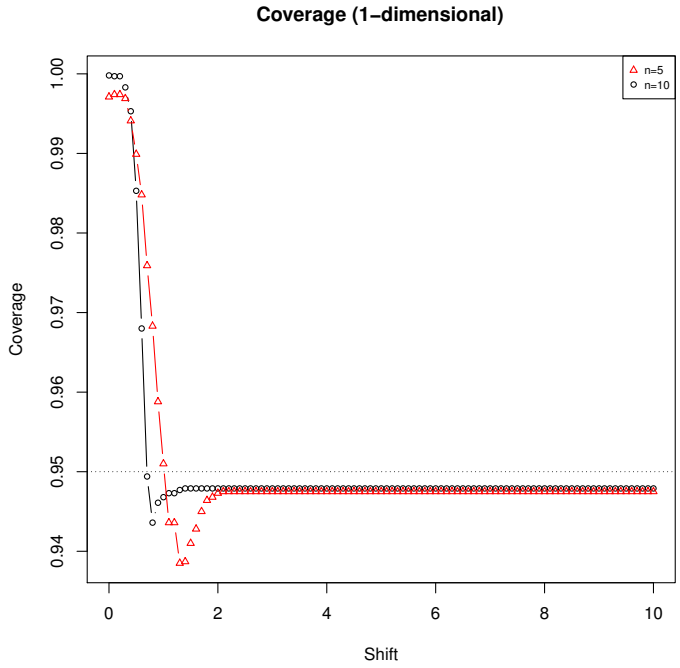

Abbildung 7.1.: Univariate Überdeckungsraten

Rote Dreiecke kennzeichnen Stichprobengrößen von 5 und schwarze Kreise von 10. Die gepunktete horizontale Linie bei 0.95 kennzeichnet das nominelle Niveau. Die Simulationen jedes Punkts wurde mit demselben Random-Seed initialisiert.

- $> 95\%$ bei Verschiebungen um weniger als 1 σ .
- $> 93.5\%$ (94%) bei Verschiebungen um weniger als 2 σ .
- $\approx 95\%$ bei Verschiebungen um mehr als 2σ .

7.1.2. Bivariate Simulationen

Das zweidimensionale Simulationsszenario entspricht weitgehend dem eindimensionalen Fall. Es werden wieder zwei unabhängige Gruppen gleicher Größe simuliert. Beide sind normalverteilt mit identischer Kovarianzstruktur $Cov(X) = \mathbb{I}_2$ und jeweils zwei Endpunkten. Die Mittelwerte werden zweidimensional verschoben. Die Stichprobengröße n ist 10 pro Gruppe. Die Stichprobengröße 5 wird diesmal nicht simuliert.

Das Abstandsmaß $d(a, b) := \max_{1 \leq k \leq p} (a_k - b_k)^2$ lässt sich nicht wie im eindimensionalen Fall vereinfachen. Da die Algorithmen in weiten Teilen identische Strukturen benutzen, deuten die Ergebnisse aus der eindimensionalen Simulationsstudie darauf hin, dass diese korrekt sind.

Die Ergebnisse sind in Abbildung 7.2 dargestellt.

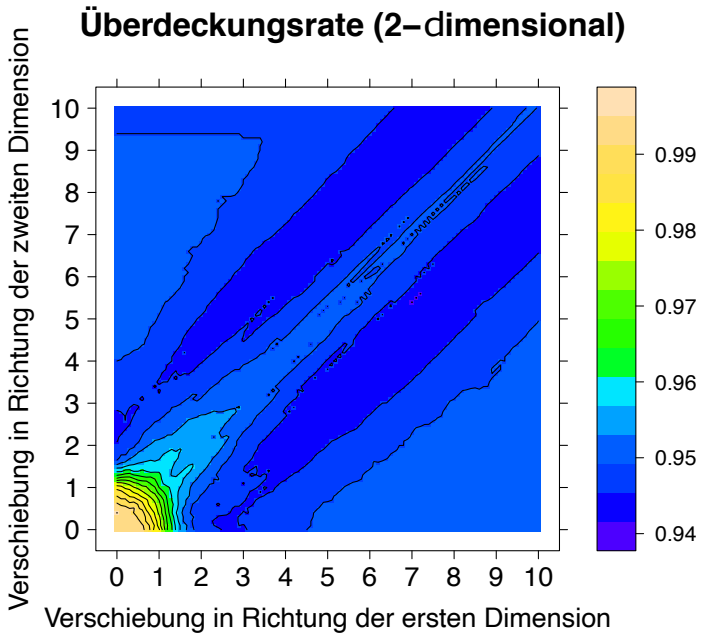

Abbildung 7.2.: Bivariate Überdeckung

Die Simulationen jedes Punkts wurde mit demselben Random-Seed initialisiert. Die Überdeckungsrate ist farbkodiert.

7.1.3. Höherdimensionale Simulationen

Höherdimensionale Simulationsergebnisse sind schwerer zu visualisieren und aus kombinatorischen Gründen nicht so umfassend durchführbar, wie es für die niedrigeren Dimensionen möglich war.

Die Ergebnisse deuten an, dass die Konfidenzintervalle mit zunehmender Dimension konservativer werden. Bei moderaten Änderungen der Korrelationsstruktur waren keine Auffälligkeiten zu beobachten. Ebenfalls waren keine Auffälligkeiten bei unterschiedlichen Korrelationsmatrizen zwischen den Gruppen oder unterschiedlichen Gruppenstichprobengrößen zu beobachten.

7.2. Power

Die Power hängt bei Äquivalenztests sehr von der Wahl der Äquivalenzgrenzen ab. Daher lässt sich ein Power-Wert beliebig erzeugen. Wenn die Grenzen gegen 0 gehen, geht auch die Power gegen 0, und wenn sie gegen ∞ gehen, geht die Power gegen 1.

Ein Vergleich zweier Verfahren ist kaum möglich, wenn die Verfahren für konzeptionell unterschiedliche Abstände schätzen. Dementsprechend sind dann auch die Äquivalenzschranken unterschiedlich zu wählen und haben in der Regel trotzdem noch verschiedene Bedeutungen.

In diesem Abschnitt wird der Versuch unternommen, das Abstandsverfahren gegen einen multiplen univarianten Äquivalenztest für die Alternative $\mu_x = \mu_y$, zu vergleichen, der in jedem Endpunkt einzeln bestanden werden muss, um insgesamt signifikant zu werden. Anstatt die Tests selbst miteinander zu vergleichen, wird die Veränderung der Power bei zunehmender Merkmalsanzahl und gleichbleibenden Äquivalenzschranken betrachtet.

Bemerkung 24:
Das angewendete Prinzip hat deutliche Schwächen beim Vergleich der Güte zweier hochdimensionaler Verfahren. Würde man beispielsweise den verwendeten Abstand in der Abstandsmethode durch denselben Abstand skaliert mit m^{-1} ersetzen, mit m als Zahl der Endpunkte, dann ergeben sich vollkommen andere asymptotische Entwicklungen, obwohl die Verfahren identisch sind, wenn nur ein Endpunkt vorliegt. Aus Mangel an guten Alternativen wird hier darauf zurückgegriffen.

Der Vorgang lässt sich folgendermaßen strukturiert darstellen:

1. Benutze dieselben Daten für beide Tests.
2. Beginne univariat.
3. Wähle für die Abstandsmethode die Stichprobengröße n und die Äquivalenzschranke so, dass eine Power von 80% erzielt wird.
4. Benutze beim multiplen Verfahren dasselbe n und passe die Äquivalenzschranken (symmetrisch) so an, dass ebenfalls 80% Power erzielt wird.
5. Verdoppele die Variablenanzahl.
6. Wähle für die Abstandsmethode n (bei unveränderter Äquivalenzschranke) so, dass eine Power von 80% erzielt wird.
7. Benutze beim multiplen Verfahren dasselbe n und nehme dieselben unveränderten Äquivalenzschranken aus Schritt 4 auch für alle neu hinzugekommenen Variablen.
8. Weiter mit Schritt 5 oder stoppen.

Das Simulationsszenario besteht wieder aus zwei unabhängigen Gruppen. Beide sind normalverteilt. Beide haben die Kovarianzstruktur $Cov(X) = \mathbb{I}$. Die Mittelwerte der Gruppen sind identisch: $\mu_x = \mu_y$. Die Gruppen sind gleich groß.

Als Gruppengrößen haben sich $18, 26, 28, 30, 32, 34, 36, 39, 41, 44, 47, 50$ und 53 ergeben. 18 bei $2^0 = 1$ Endpunkten und 53 bei $2^{12} = 4096$ Endpunkten. Siehe Abbildung 7.3.

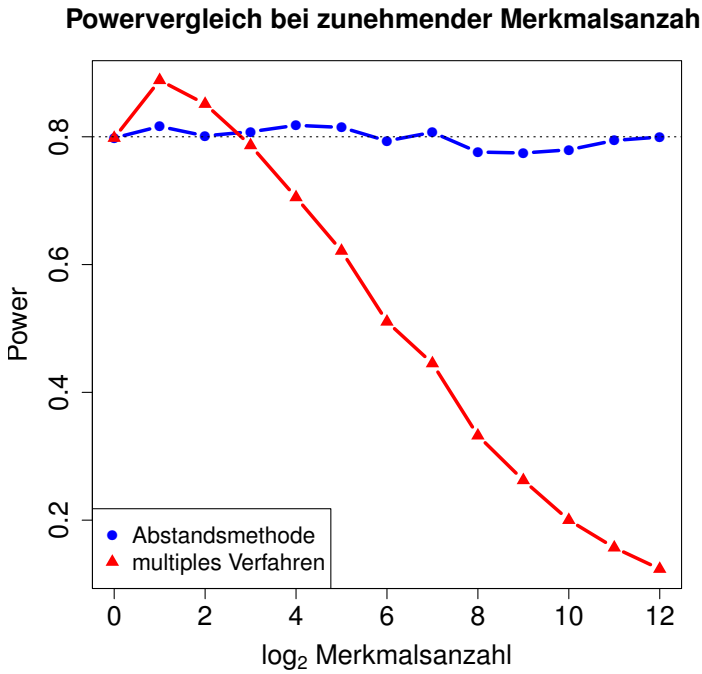

Abbildung 7.3.: Power

Die x-Achse ist zweier-logarithmisch aufgetragen von 1 bis 4096. Die zugehörigen Gruppengrößen sind: 18, 26, 28, 30, 32, 34, 36, 39, 41, 44, 47, 50 und 53. Ab 30 Variablen liegt die Güte des multiplen Verfahrens dauerhaft unter der der Abstandsmethode. Die nominelle Güte von 0.8 ist durch eine gestichelte Referenzlinie gekennzeichnet.

7.3. Fazit

Multiple Äquivalenztests werden in höheren Dimensionen praktisch unbrauchbar. Abstandsbasierte Tests können eine Alternative zu diesen darstellen. Die vergleichsweise gute Power resultiert wahrscheinlich aus einer Abschwächung der Äquivalenzforderung, die nicht mehr für jeden einzelnen Endpunkt erfüllt sein braucht. Bei sehr kleinen Verschiebungen sind die Konfidenzintervalle auch sehr konservativ. In der Regel sind das aber die Fälle, bei denen am meisten Spielraum für den positiven Äquivalenznachweis zur Verfügung steht. Insgesamt ergibt sich das Bild eines brauchbaren Verfahrens für Äquivalenztests.

Teil III.

Verfahren bei phylogenetischer Zusatzinformation

8. Multiple Testverfahren in phylogenetischem Kontext

Statistische Testprozeduren für Abundanzdaten nutzen phylogenetische Information üblicherweise, um OTUs zu definieren, die danach einfach als unterschiedliche Variablen betrachtet werden, deren relative Häufigkeiten untersucht werden.

Verfahren könnten aber aus der phylogenetische Information auch Beziehungen zwischen OTUs ableiten. Bekannt sind hier insbesondere der gewichtete und der ungewichtete UniFrac-Abstand (Definitionen 4.3 und 4.2). Multiple Testverfahren sind mir, ausser den hier entwickelten, nicht bekannt. Die Struktur phylogenetischer Daten wird in Anhang B und C diskutiert. Auf eine grundlegende hierarchische Kodierung wird im vorliegenden Kapitel aber kurz eingegangen.

Als weitere Ausgangsbasis für die Verfahren dienen unadjustierte p-Werte zu den einzelnen OTUs, die aus allgemeinen statistischen Modellen stammen können.

Nicht-zugeordnete OTUs: Eine taxonomische Baumstruktur kann durch die Namen der OTUs im Datensatz kodiert werden. Dabei wird ein eindeutiges Trennungszeichen verwendet, um die Zuteilung auf den einzelnen Ebenen voneinander zu trennen. Als Trennungszeichen wird in diesem Kapitel "." verwendet. Die Ebenen werden von links, mit der Wurzel beginnend, aufgelistet.

Beispiel 9. Eine OTU, deren Name mit "Fungi.Ascomycota.Sordariomycetes" beginnt, wäre auf der obersten taxonomischen Ebene (Königreich) dem Taxon "Fungi", auf der zweitobersten (Stamm), dem Taxon "Ascomycota" und auf der drittobersten (Klasse) dem Taxon "Sordariomycetes" zugeordnet. Beginnt ein OTU-Namen mit "Fungi.Basidiomycota.Tremellomycetes" wären das Königreich "Fungi", der Stamm "Basidiomycota" und die Klasse "Tremellomycetes".

In den vorgestellten Verfahren wird auf die taxonomische Information anstatt auf die zugrundeliegende phylogenetische zurückgegriffen. Die Verfahren wurden für die Situation entworfen und so implementiert, dass alle OTUs auf jeder taxonomischen Ebene des betrachteten

Baums einem Taxon zugeordnet sind. In der Praxis lässt sich momentan meist ein zweistelliger prozentualer Anteil der OTUs nicht auf jeder Ebene zuordnen. Es gibt grundsätzlich drei Möglichkeiten, dieses Problem zu umgehen:

- Betroffene OTUs können vor der Analyse aus dem Datensatz entfernt werden.
- Es kann jede OTU in einer Ebene, in der sie nicht zugeordnet ist, in ein künstliches Taxon "unbekannt" dieser Ebene eingeordnet werden. Jede Ebene kann mehrere dieser künstlichen Taxa haben, wenn diese verschiedene Väter haben. Wäre die erste OTU in Beispiel 9 in der zweithöchsten Ebene nicht zuordnenbar gewesen, wäre ihr Name "Fungi.unbekannt.Sordariomycetes". Wäre die zweite OTU nur in der höchsten Ebene zuordnenbar, wäre ihr Name "Fungi.unbekannt.unbekannt". Bis einschließlich der zweithöchsten Ebene wäre Sie dann zusammen mit der ersten OTU eingeordnet.
- Die unterste Ebene, für die ein entsprechendes OTU nicht zugeordnet ist, habe den Index i. Dann kopiere die Taxonbezeichnung der Ebene $i-1$ dieser OTU an die Stelle der Ebene i. Die Methode kann ggf. wiederholt werden, bis die OTU auf jeder Ebene zugeordnet ist. Bei Kodierung über OTU-Namen ist zu beachten, dass entsprechende Namen auch natürlich vorkommen können, und es sich daher empfiehlt, die Namen zusätzlich als künstlich zu kennzeichnen, um sie später eventuell besonders behandeln zu können. Beispielsweise wäre ein geeigneter Name der ersten OTU aus Beispiel 9 in der Situation die im letzten Punkt beschrieben wurde:
  "Fungi.Sordariomycetes-künstlich.Sordariomycetes". Für die zweite OTU:
  "Fungi.Cystofilobasidiales-künstlich-künstlich.Cystofilobasidiales-künstlich", wenn ihr Taxon auf der folgenden Ebene "Cystofilobasidiales" wäre. Die OTU wäre nur im Königreich zusammen mit der ersten OTU zugeordnet.

Jede dieser Methoden hat ihre Nachteile. Es werden vorhandene Informationen ignoriert, unzusammenhängende Variablen zusammengefügt oder weniger-relevante Variablen überbetont. Varianten und Mischformen dieser Methoden sind möglich, wobei letztere durch die entgegengesetzte Richtung in der Namensfindung komplizierter werden. Für die Anwendung auf reale Daten wurde im vorliegenden Text die erste Methode verwendet.

8.1. Sequentielles Testen mit datenabhängig geordneten Hypothesen mit zusätzlicher Berücksichtigung phylogenetischer Information

Die Anordnung univariater Tests anhand ihrer Varianz im datenabhängigen sequentiellen Verfahren (Seite 9) verliert ihre Vorzüge, wenn sich Variablen sehr in ihrer Skalierung unterscheiden oder sich die Messqualität zwischen ihnen unterscheidet. Es hilft nicht, die Daten zu studentisieren, da anschließend die Varianzen identisch sind und nicht mehr sinnvoll zum Anordnen genutzt werden können.

OTU-Häufigkeiten einer Probe haben in der Regel deutlich unterschiedliche Skalen (siehe Abschnitt B.3.1). Das liegt zum einen an den ökologischen Bedingungen, an die unterschiedliche Organismen unterschiedlich gut angepasst sind (Abschnitte B.3.1, B.3.1.1 und B.3.3), und an der Entwicklungsgeschichte des Lebensraums (Abschnitt B.3.2). Zum anderen hängen die Messerfolge und Messgenauigkeiten von DNA-Sequenzen aus mehreren Gründen von der jeweiligen Sequenz ab (Abschnitt C.2.5). Außerdem unterscheidet sich bei den aktuell beliebten Markergenen die Anzahl, mit der sie im Organismus vorkommen, zwischen den Organismustypen (Abschnitt C.2.5, Seite 163), während die Beziehung zwischen Umwelt und Organismus kaum von der Anzahl des Markergens abhängt (Abschnitte B.3.4 und B.1.9.1). "Marker" bezeichnet in diesem Zusammenhang ein Gen, das zur Unterscheidung zwischen verschiedenen Organismen benutzt wird und "Markersequenz" dessen Basenabfolge.

Die Sequenzen verzerren demnach die nutzbare Information für das datenabhängige sequentielle Verfahren. Andererseits sind die Sequenzen bekannt und können somit theoretisch verwendet werden, um einen Teil dieser Verzerrungen zu korrigieren. In vielen anderen Problemen, können Verzerrungen durch Stratifikation verringert werden. Im vorliegenden Fall müssten die Variablen stratifiziert werden. Wenn die Variablen anhand ihrer Sequenzähnlichkeiten in mehrere Blöcke eingeteilt werden, für die geringere nicht-versuchsspezifische Unterschiede in den Varianzen zu erwarten sind, verbessert sich die Informationsnutzung in dem Verfahren innerhalb jeder dieser Blöcke.

Es ist anzunehmen, dass sehr ähnliche Organismen ähnlich an Umweltreize (wie Versuchsbedingungen) angepasst sind, aber gegenseitig in Konkurrenz stehen können (Abschnitt B.3.1.1). Dies führt zu unklaren Erwartungen an die Korrelationen der Variablen. In den Extremfällen können zwei Organismentypen mit steigendem Umwelteinfluss gleichmäßig in ihrer Häufigkeit ansteigen. Falls sich die Organismen gegenseitig verdrängen und nur jeweils einer der beiden Typen in jeder Probe vorkommt, wären die Variablen negativ korreliert. Dies macht die Ausnutzung von Korrelationen schwierig. Die Versuchsbedingungen würden sich jedoch trotzdem auf beide absoluten Varianzen gleich auswirken.

8.1.1. Verfahren

Die Notation sei wie in Abschnitt 3.1.4. Das Verfahren kann auf folgende Weise erfolgversprechend umgesetzt werden:

1. Wähle eine taxonomische Ebene Q_i oder bilde auf eine andere Art Cluster der Variablen $X_j, j \in Q_1$ anhand ihrer Sequenzähnlichkeiten. (Q_1 ist gemäß Definition 3.3 die Menge der Indizes der Blätter.)
2. Ermittle die unadjustierten univariaten p-Werte p_j für alle OTUs $j \in Q_1$.
3. Adjustiere die p-Werte zunächst mittels Bonferroni für die Anzahl der Cluster $\#Q_i$ durch $\tilde{p}_j = \min(1, p_j \cdot \#Q_i)$. Alternativ kann das gewichtete Bonferroni-Verfahren verwendet werden, wobei sich die Größer der Cluster (d.h. Anzahl enthaltener Blätter) als Grundlage der Gewichte anbietet.

4. Für jeden Cluster $q \in Q_i$: Sortiere die Variablen $X_j, j \in Q_i$ nach absteigender absoluter Varianz. Sei durch l_q eine Indizierung aller Elemente von q definiert, mit der Eigenschaft:
$$\text{Var}(X_{l_q(1)}) \geq \ldots \geq \text{Var}(X_{l_q(\#q)}) \,.$$

5. Definiere den vollständig adjustierten p-Wert $\tilde{p}'_j$ jeder Variable X_j durch das Maximum aller Cluster-adjustierten p-Werte, die bis zu dieser Stelle in dem Cluster stehen:
$$\forall_{q \in Q_i} \forall_{s \in \{1,\ldots,\#q\}} : \tilde{p}'_{l_q(s)} = \max_{t \leq s} (\tilde{p}_{l_q(t)}) \,.$$

Die vollständig adjustierten p-Werte liefern den endgültigen Testentscheid unter Einhaltung der FWER, da das sequentielle Verfahren jeweils das Bonferroni-adjustierte Niveau einhält und das Bonferroni-Verfahren für die Anzahl der Cluster korrigiert. Welche der beiden Adjustierungen zuerst ausgeführt wird, hat keinen Einfluss auf die endgültigen p-Werte.

Beispiel 10. In Antweiler *et al.* (2017) wurde das Mikrobiom von Bodenproben aus 2 Standorten 3 Bodentypen unter Kontroll- und Behandlungsbedingung untersucht. Hier soll nun der p-Wert für den Bodentyp aus diesem 3-faktoriellen gewöhnlichen linearen Modell betrachtet und durch das Verfahren aus Abschnitt 8.1 adjustiert werden. OTUs, die nicht vollständig zugeordnet werden konnten, wurden entfernt. Taxonomie, Indices aus den verschiedenen Ebenen, Varianzen und unadjustierte p-Werte sind in Tabelle 8.1 dargestellt. Von der 7 stufigen Taxonomie wurden hier, um die Tabellen darstellbar zu halten, nur die ersten 3 (Königreich, Stamm und Ordnung) verwendet. Die relativen Häufigkeiten der OTUs, deren Taxonomie auf den ersten 3 Ebenen übereinstimmten, wurden hierfür zusammengezählt. Die Zeilen der Tabelle sind lexikographisch sortiert nach Q_1, Q_2 und absteigender Varianz. Diese Reihenfolge bleibt in allen Tabellen (8.1, 8.2, 8.3, 8.4) dieses Beispiels identisch.

Für die Cluster-Bildung wurde Q_2 gewählt. Damit ergeben sich die 4 Cluster: $\{1, \ldots, 9\}$, $\{10, \ldots, 13\}$, $\{14\}$ und $\{15\} \in Q_2$. Die Cluster-Adjustierten p-Werte $\tilde{p}$ ergeben sich durch Multiplikation der Zahl 4 mit den unadjustierten p-Werten p und anschließendem Abrunden größerer Werte auf 1.0 (Tabelle 8.2). Wären die Tabellen noch nicht innerhalb der Cluster nach absteigender Varianz sortiert, müsste dies jetzt geschehen. Die vollständig adjustierten p-Werte $\tilde{p}'$ werden berechnet, indem, von oben beginnend, jeder p-Wert $\tilde{p}$ auf den über ihm stehenden Wert desselben Clusters angehoben wird, sofern dieses möglich ist.

Ein analoges Vorgehen mit einem Blätter-gewichteten Bonferroni-Schritt ist in Tabelle 8.3 dargestellt. Die p-Werte $\tilde{p}$ werden hier durch Multiplikation mit dem Quotienten ("Anzahl Blätter insgesamt" / "Anzahl Blätter im Cluster") berechnet. Die anderen Schritte sind identisch.

Die Ergebnisse sind nochmal in Tabelle 8.4 zusammengefasst. Die Tabelle enthält zusätzlich die p-Werte p', die das ursprüngliche datenabhängig-geordnete Verfahren (Seite 9)

liefert. Das sind dieselben Werte, die bei der Wahl der Wurzel-Ebene Q_3 als Cluster-Menge herauskommen würden. In schwarzem Fettdruck wurden p-Werte hervorgehoben, die unter 0.05 liegen und in dunkelgrauem schrägen Fettdruck, die unter 0.1 .

Die Verfahren führen in dem Beispiel zu ähnlichen Ergebnissen. Im ersten Cluster verhindert die Variable mit höchster Varianz niedrige p-Werte für alle Verfahren. Im zweiten Cluster wird die signifikante Variable von allen Verfahren gefunden, wobei sich die Höhe des p-Werts deutlich unterscheidet. Die vollständigen Adjustierungen profitieren hier davon, dass die Varianzen der restlichen Variablen des Clusters vergleichbar sind und nur die signifikante Variable sich davon abhebt. In den beiden letzten Clustern fand für die vollständigen Adjustierungen effektiv keine Sortierung statt. Das ursprüngliche Verfahren hat die Signifikanz des letzten Clusters gefunden. Das Cluster-Anzahl-gewichtete Verfahren kommt in die Nähe.

Die Menge der Blätter war in diesem Beispiel aus praktischen Gründen viel geringer als es in realen Auswertungen der Fall sein würde. Die Art der phylogenetischen Zusammenhänge hängt darüber hinaus von der betrachteten taxonomischen Ebene ab – siehe Abschnitt C.3.4 auf Seite 169 und B.3.4 auf Seite 147.

Bemerkung 25 (zu Beispiel 10):
Die Hypothesen zu den einzelnen Variablen sind nicht bekannt. Heuristisch gesehen könnte man sagen, dass die Variablen aus so viele Variablen zusammengesetzt sind (siehe Abbildung 8.1 auf Seite 84), dass jeweils die Alternative gelten müsse. Das gälte aber höchstens für die ersten zwei (eventuell drei) Cluster des Beispiels. Für eine solche Heuristik wäre die Art der Zusammenfassung der Variablen sehr grob. (Umgekehrt kann man im Allgemeinen oft alle Variablen soweit aufspalten, dass irgendwo ein Effekt zu vermuten ist.) Es lässt sich auch vermuten, dass Fehler erster Art bei niedrigem Niveau nur selten auftreten, wenn die Verfahren ihr Niveau einhalten und für Power-Betrachtungen die Anzahl kleiner p-Werte gezählt werden könnte. Die Voraussetzungen unter denen die, für die unadjustierten p-Werte verwendeten, gewöhnlichen linearen Modelle ihr Niveau einhalten, sind aber nicht strikt erfüllt.

8.2. Hierarchisches Testen auf phylogenetisch motivierten Clustern

Phylogenetische Information lässt sich zur Definition der Cluster eines hierarchischen Testverfahrens (Abschnitt 3.1.4) verwenden. Die vollständige Baumstruktur einer gewöhnlichen taxonomischen Analyse bietet sich hierzu an. Diese hat darüber hinaus den Vorteil, dass alle Analyseergebnisse zwangsläufig, verglichen mit üblichen multiplen Analysen, kompakt und verständlich interpretier- und präsentierbar sind.

Tabelle 8.1.: Sequentielle Adjustierung Daten

Taxonomie	$\mathbf{q} \in \mathbf{Q}_1$	$\mathbf{q} \in \mathbf{Q}_2$	$\mathbf{q} \in \mathbf{Q}_3$	**Varianz**	**p**
Fungi.Ascomycota.Leotiomycetes	$\{1\}$	$\{1,\ldots,9\}$	$\{1,\ldots,15\}$	4.9e-04	3.0e-01
Fungi.Ascomycota.Dothideomycetes	$\{2\}$	$\{1,\ldots,9\}$	$\{1,\ldots,15\}$	1.4e-04	2.0e-04
Fungi.Ascomycota.Saccharomycetes	$\{3\}$	$\{1,\ldots,9\}$	$\{1,\ldots,15\}$	6.8e-05	2.5e-05
Fungi.Ascomycota.Pezizomycetes	$\{4\}$	$\{1,\ldots,9\}$	$\{1,\ldots,15\}$	3.7e-05	5.1e-02
Fungi.Ascomycota.Lecanoromycetes	$\{5\}$	$\{1,\ldots,9\}$	$\{1,\ldots,15\}$	5.1e-06	5.8e-02
Fungi.Ascomycota.Eurotiomycetes	$\{6\}$	$\{1,\ldots,9\}$	$\{1,\ldots,15\}$	2.5e-06	8.7e-02
Fungi.Ascomycota.Sordariomycetes	$\{7\}$	$\{1,\ldots,9\}$	$\{1,\ldots,15\}$	2.3e-06	1.1e-01
Fungi.Ascomycota.Taphrinomycetes	$\{8\}$	$\{1,\ldots,9\}$	$\{1,\ldots,15\}$	7.1e-07	8.6e-01
Fungi.Ascomycota.Orbiliomycetes	$\{9\}$	$\{1,\ldots,9\}$	$\{1,\ldots,15\}$	2.8e-07	1.2e-01
Fungi.Basidiomycota.Ustilaginomycetes	$\{10\}$	$\{10,\ldots,13\}$	$\{1,\ldots,15\}$	5.7e-04	1.3e-05
Fungi.Basidiomycota.Microbotryomycetes	$\{11\}$	$\{10,\ldots,13\}$	$\{1,\ldots,15\}$	7.0e-06	2.9e-01
Fungi.Basidiomycota.Tremellomycetes	$\{12\}$	$\{10,\ldots,13\}$	$\{1,\ldots,15\}$	6.0e-06	2.5e-01
Fungi.Basidiomycota.Agaricomycetes	$\{13\}$	$\{10,\ldots,13\}$	$\{1,\ldots,15\}$	1.1e-06	4.3e-01
Fungi.Chytridiomycota.Chytridiomycetes	$\{14\}$	$\{14\}$	$\{1,\ldots,15\}$	4.6e-04	1.4e-02
Fungi.Glomeromycota.Glomeromycetes	$\{15\}$	$\{15\}$	$\{1,\ldots,15\}$	3.1e-03	2.1e-02

Tabelle 8.2.: Sequentielle Adjustierung Cluster-gewichtet

$\mathbf{q} \in \mathbf{Q}_1$	$\mathbf{q} \in \mathbf{Q}_2$	**Varianz**	**p**	$\mathbf{n}_{\text{cluster}}$	$\tilde{\mathbf{p}}$	$\tilde{\mathbf{p}}'$
$\{1\}$	$\{1,\ldots,9\}$	4.95e-04	2.98e-01	4	1.00e+00	1.00e+00
$\{2\}$	$\{1,\ldots,9\}$	1.44e-04	1.95e-04	4	7.82e-04	1.00e+00
$\{3\}$	$\{1,\ldots,9\}$	6.78e-05	2.55e-05	4	1.02e-04	1.00e+00
$\{4\}$	$\{1,\ldots,9\}$	3.75e-05	5.10e-02	4	2.04e-01	1.00e+00
$\{5\}$	$\{1,\ldots,9\}$	5.07e-06	5.82e-02	4	2.33e-01	1.00e+00
$\{6\}$	$\{1,\ldots,9\}$	2.49e-06	8.68e-02	4	3.47e-01	1.00e+00
$\{7\}$	$\{1,\ldots,9\}$	2.28e-06	1.13e-01	4	4.51e-01	1.00e+00
$\{8\}$	$\{1,\ldots,9\}$	7.09e-07	8.56e-01	4	1.00e+00	1.00e+00
$\{9\}$	$\{1,\ldots,9\}$	2.80e-07	1.22e-01	4	4.89e-01	1.00e+00
$\{10\}$	$\{10,\ldots,13\}$	5.74e-04	1.33e-05	4	5.32e-05	5.32e-05
$\{11\}$	$\{10,\ldots,13\}$	7.00e-06	2.89e-01	4	1.00e+00	1.00e+00
$\{12\}$	$\{10,\ldots,13\}$	6.01e-06	2.48e-01	4	9.92e-01	1.00e+00
$\{13\}$	$\{10,\ldots,13\}$	1.12e-06	4.27e-01	4	1.00e+00	1.00e+00
$\{14\}$	$\{14\}$	4.62e-04	1.42e-02	4	5.66e-02	5.66e-02
$\{15\}$	$\{15\}$	3.06e-03	2.13e-02	4	8.52e-02	8.52e-02

Tabelle 8.3.: Sequentielle Adjustierung Blätter-gewichtet

$\mathbf{q}\in\mathbf{Q}_1$	$\mathbf{q}\in\mathbf{Q}_2$	**Varianz**	**p**	**Faktor**	$\tilde{\mathbf{p}}$	$\tilde{\mathbf{p}}'$
$\{1\}$	$\{1,\ldots,9\}$	4.95e-04	2.98e-01	15 / 9	4.97e-01	4.97e-01
$\{2\}$	$\{1,\ldots,9\}$	1.44e-04	1.95e-04	15 / 9	3.26e-04	4.97e-01
$\{3\}$	$\{1,\ldots,9\}$	6.78e-05	2.55e-05	15 / 9	4.25e-05	4.97e-01
$\{4\}$	$\{1,\ldots,9\}$	3.75e-05	5.10e-02	15 / 9	8.50e-02	4.97e-01
$\{5\}$	$\{1,\ldots,9\}$	5.07e-06	5.82e-02	15 / 9	9.70e-02	4.97e-01
$\{6\}$	$\{1,\ldots,9\}$	2.49e-06	8.68e-02	15 / 9	1.45e-01	4.97e-01
$\{7\}$	$\{1,\ldots,9\}$	2.28e-06	1.13e-01	15 / 9	1.88e-01	4.97e-01
$\{8\}$	$\{1,\ldots,9\}$	7.09e-07	8.56e-01	15 / 9	1.00e+00	1.00e+00
$\{9\}$	$\{1,\ldots,9\}$	2.80e-07	1.22e-01	15 / 9	2.04e-01	1.00e+00
$\{10\}$	$\{10,\ldots,13\}$	5.74e-04	1.33e-05	15 / 4	4.99e-05	4.99e-05
$\{11\}$	$\{10,\ldots,13\}$	7.00e-06	2.89e-01	15 / 4	1.00e+00	1.00e+00
$\{12\}$	$\{10,\ldots,13\}$	6.01e-06	2.48e-01	15 / 4	9.30e-01	1.00e+00
$\{13\}$	$\{10,\ldots,13\}$	1.12e-06	4.27e-01	15 / 4	1.00e+00	1.00e+00
$\{14\}$	$\{14\}$	4.62e-04	1.42e-02	15 / 1	2.12e-01	2.12e-01
$\{15\}$	$\{15\}$	3.06e-03	2.13e-02	15 / 1	3.20e-01	3.20e-01

Tabelle 8.4.: Sequentielle Adjustierungen Ergebnisse

$\mathbf{q}\in\mathbf{Q}_1$	$\mathbf{q}\in\mathbf{Q}_2$	**Varianz**	**p**	$\mathbf{p}'$	$\mathbf{p}_{\text{cluster}}$	$\mathbf{p}_{\text{leafs}}$
$\{1\}$	$\{1,\ldots,9\}$	4.95e-04	2.98e-01	2.98e-01	1.00e+00	4.97e-01
$\{2\}$	$\{1,\ldots,9\}$	1.44e-04	**1.95e-04**	2.98e-01	1.00e+00	4.97e-01
$\{3\}$	$\{1,\ldots,9\}$	6.78e-05	**2.55e-05**	2.98e-01	1.00e+00	4.97e-01
$\{4\}$	$\{1,\ldots,9\}$	3.75e-05	***5.10e-02***	2.98e-01	1.00e+00	4.97e-01
$\{5\}$	$\{1,\ldots,9\}$	5.07e-06	***5.82e-02***	2.98e-01	1.00e+00	4.97e-01
$\{6\}$	$\{1,\ldots,9\}$	2.49e-06	***8.68e-02***	2.98e-01	1.00e+00	4.97e-01
$\{7\}$	$\{1,\ldots,9\}$	2.28e-06	1.13e-01	2.98e-01	1.00e+00	4.97e-01
$\{8\}$	$\{1,\ldots,9\}$	7.09e-07	8.56e-01	8.56e-01	1.00e+00	1.00e+00
$\{9\}$	$\{1,\ldots,9\}$	2.80e-07	1.22e-01	8.56e-01	1.00e+00	1.00e+00
$\{10\}$	$\{10,\ldots,13\}$	5.74e-04	**1.33e-05**	**2.13e-02**	**5.32e-05**	**4.99e-05**
$\{11\}$	$\{10,\ldots,13\}$	7.00e-06	2.89e-01	2.98e-01	1.00e+00	1.00e+00
$\{12\}$	$\{10,\ldots,13\}$	6.01e-06	2.48e-01	2.98e-01	1.00e+00	1.00e+00
$\{13\}$	$\{10,\ldots,13\}$	1.12e-06	4.27e-01	4.27e-01	1.00e+00	1.00e+00
$\{14\}$	$\{14\}$	4.62e-04	**1.42e-02**	2.98e-01	***5.66e-02***	2.12e-01
$\{15\}$	$\{15\}$	3.06e-03	**2.13e-02**	**2.13e-02**	***8.52e-02***	3.20e-01

8.2.1. Verallgemeinerung der Hypothesenstruktur

Bemerkung 26:
Theorem 1 aus Meinshausen (2008) ist zwar für Hypothesen auf hierarchisch-angeordneten Clustern von Regressions-Koeffizienten definiert, aber Algorithmus und Beweis erfordern bei genauerer Betrachtung weniger Struktur. Insbesondere sind keine Anforderungen an die Beziehungen zwischen den Hypothesen auf verschiedenen hierarchischen Ebenen nötig. Das wird für Korollar 8.3 wichtig.

Satz 8.1 (Allgemeinere Hierarchische Adjustierung):
Seien Q und $(Q_i)_{i=1,\dots,k}$ wie in Definition 3.3 mit $k \in \mathbf{N}$ Ebenen und indizieren die Hypothesen $(H_{q,i})_{q\in Q_i, i\in\{1,\dots,k\}}$.
Dann halten die folgend adjustierten p-Werte $\tilde{p}'_{q,i} := \max_{j\geq i, r\in Q_j, q\subset r}(\tilde{p}_{r,j})$ mit $\tilde{p}_{q,j} := \min\left(1,\, p_{q,j}\cdot\left(\#\bigcup_{s\in Q} s\right)/\#q\right)$ die FWER ein.

Der Beweis folgt formal dem Beweis aus Meinshausen (2008).

Beweis von Satz 8.1. Seien Q, k, $(Q_i)_{i=1,\dots,k}$ und $(H_{q,i})_{1\leq i\leq k, q\in Q_i}$ wie in Satz 8.1 definiert. Seien für alle $1 \leq i \leq k$:

$$\begin{aligned} Q_{i,\text{rej}} &:= \{q \in Q_i | \tilde{p}'_{q,i} \leq \alpha\}, \\ Q_{i,0} &:= \{q \in Q_i | H_{q,i} \text{ ist wahr}\} \text{ und} \\ \tilde{Q}_{i,0} &:= \{q \in Q_{i,0} | \nexists_{\substack{j>i \\ r\in Q_{j,0}}} : q \subset r\}. \end{aligned} \tag{8.1}$$

Die FWER lässt sich dann folgend begrenzen:

$$\begin{aligned} &P\left(\bigcup_{i=1}^{k}\left(Q_{i,\text{rej}} \cap Q_{i,0}\right) \neq \emptyset\right) = P\left(\exists_{\substack{q\in Q_{i,0} \\ 1\leq i\leq k}} : \tilde{p}'_{q,i} \leq \alpha\right) \\ =&P\left(\exists_{\substack{q\in \tilde{Q}_{i,0} \\ 1\leq i\leq k}} : \tilde{p}'_{q,i} \leq \alpha\right) \leq P\left(\exists_{\substack{q\in \tilde{Q}_{i,0} \\ 1\leq i\leq k}} : \tilde{p}_{q,i} \leq \alpha\right) \\ \leq&\sum_{i=1}^{k}\sum_{q\in\tilde{Q}_{i,0}} P(\tilde{p}_{q,i} \leq \alpha) \leq \sum_{i=1}^{k}\sum_{q\in\tilde{Q}_{i,0}} \alpha\frac{\#q}{\#\bigcup_{r\in Q} r}\,. \end{aligned}$$

Es reicht zu zeigen, dass $\sum_{i=1}^{k}\sum_{q\in\tilde{Q}_{i,0}} \#q \leq \#\bigcup_{r\in Q} r$.
Für alle $i, j \in \{1,\dots,k\}$, $q \in \tilde{Q}_{i,0}$, $r \in \tilde{Q}_{j,0}$, $q \neq r$ gilt nach Definition von $\tilde{Q}_{i,0}$ aus Gleichung (8.1) und des Ebenen-Begriffs aus Definition 3.3: $q \cap r = \emptyset$.
Weiter gilt: $\bigcup_{\substack{q\in\tilde{Q}_{i,0} \\ 1\leq i\leq k}} q \subset \bigcup_{r\in Q}$.

Daher gilt $\sum_{i=1}^{k}\sum_{q\in\tilde{Q}_{i,0}} \#q \leq \#\bigcup_{r\in Q} r$, womit der Beweis abgeschlossen ist. □

8.2.2. Taxonomische Adjustierung

Bei der Kommunikation der Ergebnisse hierarchischer Verfahren im taxonomischen Kontext können Missverständnisse entstehen. Wenn gesagt wird, dass die "Familie xyz" signifikant war, kann dies so verstanden werden, als ob sich die Häufigkeiten aller Organismen dieser Familie zusammengezählt zwischen einer Versuchsgruppe und einer Kontrollgruppe unterscheiden. Das ist im allgemeinen jedoch nicht nötig. Beispielsweise könnte Gattung 1 dieser Familie nur in der Versuchsgruppe auftauchen und Gattung 2 nur in der Kontrollgruppe aber mit derselben Häufigkeit, so dass sich die Häufigkeiten der Familie deshalb nicht zwischen den Gruppen unterscheiden würde. Idealerweise wird dies in Ergebnissen angesprochen.

Wie im Folgenden gezeigt wird, kann alternativ ein Test verwendet werden, bei dem diese engere Interpretation der Ungleichheit von taxonomischen Gruppen korrekt ist und der Fehler erster Art durch das hierarchische Schema aus Satz 8.1 eingehalten wird. Bei diesem Schema wird die Idee eines Fokus-Level aus Goeman & Mansmann (2008) übernommen. Dieser Level stellt hier eine Ebene i in einem hierarchischen Schema dar, auf der mit dem Testen begonnen wird. Im Gegensatz zu Goeman & Mansmann (2008) werden keine Aussagen über die Cluster oberhalb dieser Ebene i getroffen. Betrachtet man die Knoten $r \in Q_i$ der Ebene i, dann haben die Unterbäume $(q \in Q_j)_{j \leq i, q \subset r}$ jedes Knotens r dieser Ebene wieder die vollständige hierarchische Test-Struktur für Satz 8.1. Zusammen mit einer Adjustierung für das Testen mehrerer Bäume, ergibt sich eine geeignete Test-Prozedur. Das Verfahren wird zunächst in Satz 8.2 allgemein definiert und dann in Korollar 8.3 auf die eben beschriebene Situation ausgerichtet.

Satz 8.2 (allgemeine taxonomische Adjustierung):
Seien $Q, (Q_i)_{i=1,\dots,k}$ wie in Definition 3.3 mit $k \in \mathbf{N}$ Ebenen und indizieren die Hypothesen $(H_{q,i})_{q \in Q_i, i \in \{1,\dots,k\}}$.
Sei i die Ebene, ab der getestet werden soll. Für jedes $j \leq i$ und $q \in Q_j$ sei $p_{q,j}$ der unadjustierte p-Wert. Diese werden durch das Verfahren aus Abschnitt 3.1.1 adjustiert. Für jedes $q \in Q_i$, $j \leq i$, und $s \in Q_j, s \subset q$ sei $\tilde{p}_{s,j} = \min(1, p_{s,j} \cdot \#Q_i)$ der so adjustierte p-Wert.
Auf jedes $q \in Q_i$ mit seiner Unterbaumstruktur und den zugehörigen p-Werten $\tilde{p}_{s,j}$, $j \leq i$ wird das Verfahren aus Satz 8.1 angewendet.
Für jedes $j \leq i$ und $q \in Q_j$ sei $\tilde{p}'_{q,j} = \max_{j \leq l \leq i, r \in Q_l, q \subset r}(\tilde{p}_{r,l})$ der so adjustierte p-Wert.
Die so definierten p-Werte $\tilde{p}'$ halten die FWER ein.

Beweis von Satz 8.2. Seien Q, $(H_{q,j})_{j \in \{1,\dots,i\}, q \in Q_j}$, i, p und $\tilde{p}$ wie im Satz.
Für alle $q \in Q_i : \tilde{p}'_q = 1$ ist nichts zu zeigen. OBdA gelte für alle $q \in Q_i : \tilde{p}'_q < 1$ und daher auch $\tilde{p}_{q,i} < 1$.
Für alle $q \in Q_i$ seien $b_{q,i} := \tilde{p}_{q,i}/p_{q,i} = \#Q_i$ die Bonferroni-Adjustierungsfaktoren.
Für alle $j \leq i$ und $s \in Q_j$ mit $q \in Q_i$ und $s \subset q$ sei $p'_{s,j} := \tilde{p}'_{s,j}/b_{q,i} = \max_{j \leq l \leq i, r \in Q_l, s \subset r}(\tilde{p}_{r,l})/b_{q,i}$ der nur nach Satz 8.1 adjustierte p-Wert. Zum Niveau α halten diese p' den Fehler erster Art,

für den jeweiligen Unterbaum als Familie, ein. Daher halten die $\tilde{p}'$ diesen Fehler für die einzelnen Unterbäume mit Wurzel $q \in Q_i$ zum jeweiligen Niveau $\alpha/b_{q,i}$ ein. Die FWER ist demnach durch $\alpha \sum_{q \in Q_i} 1/b_{q,i} \leq \alpha$ begrenzt. □

Aus Satz 8.2 folgt nun unmittelbar das für die Anwendung benötigte:

Korollar 8.3 (Korollar zu Satz 8.2):
Sei Q eine Indexmenge taxonomischer Einheiten auf k Ebenen mit der hierarchischen Anordnung $(Q_i)_{1 \leq i \leq k}$. Für jede taxonomische Einheit $q \in Q_i$, $i \in \{1, \ldots, k\}$ sei $p_{q,i}$ der p-Wert eines nicht-liberalen univariaten Tests auf der relativen Häufigkeit dieser Einheit in den Proben. Dann funktioniert der hierarchische Algorithmus aus Satz 8.2.

Bemerkung 27:
Die zu testende Variable X_q in jedem Knoten $q \in Q$ bildet sich aus der Summe der Werte der Blätter, die Nachfahren des Knotens sind - bzw. aus dem Blatt selbst, wenn er ein Blatt ist:

$$\forall_{q \in Q} : X_q := \sum_{i \in q} X_i \,.$$

Falls die Daten transformiert analysiert werden sollen, muss diese Transformation f nach der Addition stattfinden:

$$\forall_{q \in Q} : X'_q := f(X_q) = f\Big(\sum_{i \in q} X_i \Big) \,.$$

Die Hypothesen oberhalb der Blätter lassen sich nicht aus den Hypothesen der Blätter konstruieren. Die Eigenschaft, dass, wenn eine Kind-Hypothese falsch ist, auch die Hypothese des Vaters falsch sein muss, die hier im allgemeinen nicht erfüllt wäre, ist für Meinshausens Algorithmus zwar nicht notwendig, aber bei der Fokus-Level-Methode (Goeman & Mansmann (2008)) für die Ebenen über dem Fokus-Level.

Da relative Daten getestet werden, ist auf einem vollständigen taxonomischen Baum die Wurzel-Hypothese wahr. Daher muss auf einer niedrigeren Ebene mit dem Algorithmus begonnen werden. Rückschlüsse auf höhere Ebenen sind im allgemeinen nicht möglich. Die Methode unterscheidet sich durch ihren Hypothesenwechsel zwischen den Ebenen sowohl von der Fokus-Level-Methode als auch vom Meinshausens-Verfahren (Meinshausen (2008)). Zur Unterscheidbarkeit bezeichne ich die Methode als "taxonomische Adjustierung".

Bemerkung 28:
Wie sinnvoll der Hypothesenwechsel ist, hängt auch davon ab, in wie weit die Häufigkeiten von OTU-Reads denen der Organismen entsprechen (Abschnitt C.2.5) und hängt von der Wahl des Sequenzierungsbereichs ab (Abschnitt C.2). Insbesondere durch mehrfach-vorkommende Markergene (Abschnitte C.2.5.1-C.2.5.3) und sequenzabhängige Ausbeute in PCR-Verfahren (Abschnitte C.2.5.7 und C.2.5.4) ist dies in den meisten aktuellen Analysen zu bedenken. Die gemessene Häufigkeit eines Taxons kann sich in solchen Situationen, bei gleichbleibender realer Häufigkeit, durch ihre Zusammensetzung aus verschiedenen Organismen ändern.

Bemerkung 29:
Den Knoten über der Ebene, auf der der Test begonnen wird, lassen sich auch Hypothesen zuweisen, die anders konstruiert werden als die darunterliegenden Hypothesen. Wenn sie als Schnitthypothesen derer der Anfangsebene gebildet werden, funktioniert, von dieser Ebene aufwärts, der Bottom-Up-Algorithmus aus Goeman & Mansmann (2008). Das erschwert allerdings die Kommunikation der Testergebnisse.

Beispiel 11 (Beispiel zu Korollar 8.3). Für ein Mikrobiom-Experiment gelte das statistische Modell:

$$f(Y) = DB + \varepsilon\,. \tag{8.2}$$

$Y \in \mathbf{R}^{n\times m}$ sei die Zufallsmatrix der gemessenen relativen Häufigkeiten von m OTUs in n Proben.
$D \in \mathbf{R}^{n\times d}$ sei die Design-Matrix des Experiments, mit d Spalten.
B sei die Matrix der Koeffizienten.
ε sei die Zufallsmatrix der Residuen, deren i.i.d. Zeilen multivariat-Normalverteil mit Erwartungswert $(0,\dots,0)$ seien.
f sei die Identitätsfunktion.

Sei $Q = \{q_1,\dots,q_{m'}\}$ eine vollständige Hierarchie mit k Ebenen $(Q_l)_{l=1,\dots,k}$ und m' Knoten.
Sei $C \in \{0,1\}^{m\times m'}$, mit

$$\forall_{\substack{1\le i\le m\\ 1\le j\le m'}} : C_{ij} = 1 :\Leftrightarrow j \in q_i\,.$$

Dann beschreibt

$$f(YC) = DBC + \varepsilon C$$

die statistischen Modelle des Experiments für das gesamte Korollar 8.3 mit den Hypothesen $(f(YC_{.i}) \sim H_i)_{i=1,\dots,m'}$. Da f hier linear ist, können Schätzungen für das B und ε aus Gleichung (8.2) verwendet werden – denn für jeden Spaltenvektor c aus C gilt:

$$\begin{aligned}
\beta :=& \text{argmin}_{\beta\in\mathbf{R}^d}\big((Yc - D\beta)^T(Yc - D\beta)\big)\\
\hat{B}_{.i} :=& \text{argmin}_{\hat{B}_{.i}\in\mathbf{R}^d}\big((Y_{.i} - D\hat{B}_{.i})^T(Y_{.i} - D\hat{B}_{.i})\big) \text{ für alle } i = 1,\dots,m\\
\Rightarrow& (Yc - D\beta)^T(Yc - D\beta) = \big((I - D(D^TD)^{-1}D^T)Yc\big)^T\big((I - D(D^TD)^{-1}D^T)Yc\big)\\
=& (Yc - D\hat{B}c)^T(Yc - D\hat{B}c)\,.
\end{aligned}$$

Wenn $\tilde{B}$ die Parameter des Modells unter den Nullhypothesen, ν, $\tilde{\nu}$ die Freiheitsgrade des Modelles bzw. des Null-Modells und F die Verteilungsfunktion der F-Verteilung bezeichnen, lassen sich die unadjustierten p-Werte der linearen Modelle für jeden Spaltenvektor c aus C durch

$$p_c := 1 - F_{\nu-\tilde{\nu},n-\nu}\left(\frac{c^T(\hat{B}-\hat{\tilde{B}})^TD^TD(\hat{B}-\hat{\tilde{B}})c/(\nu-\tilde{\nu})}{c^T(Y - D\hat{B})^T(Y - D\hat{B})c/(n-\nu)}\right)$$

berechnen. Die taxonomische Adjustierung lässt sich zusammengefasst schreiben als:

$$\forall_{1\leq i\leq m'} \quad \tilde{p}'_{C.i} := \min\left(1, \max_{\substack{1\leq j\leq m' \\ C_{.i}^T(C_{.i}-C_{.j})=0}} \left(p_{C.j}\frac{m}{C_{.j}^T C_{.j}}\right)\right) .$$

8.2.3. Visualisierung

Die Ergebnisse hierarchischer Tests auf taxonomischen Strukturen lassen sich, selbst bei einer großen Anzahl an signifikanten Variablen, in einer einzigen Grafik übersichtlich zusammenfassen. Dabei repräsentiert die Größe der Darstellung eines Taxons ihren Anteil in den Versuchsdaten. Damit auch die Ergebnisse in kleineren Bereichen erkennbar bleiben, wird eine Vektorgrafik erzeugt, die interaktiv vergrößerbar ist (Abb. 8.1).

In der Prozedur zur Herstellung dieser Grafiken übertrage ich das aus fMRI-Analysen bekannte Konzept des Statistical Parametric Mappings, bei dem p-Werte farbkodiert auf Gehirn-Strukturen gelegt werden, auf polar-gezeichnete taxonomische Bäume, die ungefähr geschachtelten Tortendiagrammen entsprechen.

Farbkodierung: Für die Umwandlung der p-Werte in Farbwerte gibt es viele Möglichkeiten. Eine Möglichkeit, die hier verwendet wurde: Für p-Werte wird eine Signifikanz-Grenze festgelegt. p-Werte, die darüber liegen, erhalten in einer RGB-Farbkodierung für alle drei Farbwerte 50% des Maximalwerts. Für signifikante Pixel wird der Rot-Wert auf 100% gesetzt und den Blau-Wert auf 0%. Der Grün-Wert wurde definiert als:

$$\frac{\log(1+\epsilon)-\log(p+\epsilon)}{\log(1+\epsilon)-\log(0+\epsilon)}(1-b)+b\,,$$

wobei $\epsilon := 10^{-8}$ und $b := 0.2$ gesetzt war. In Prozent ausgedrückt, kommt noch der Faktor 100% dazu. Die log-Transformation macht aus dem p-Wert (Wahrscheinlichkeit) einen Informationswert. Für die Darstellung von p-Werten wird dies häufig verwendet. Dieser Informationswert wird hier noch affin transformiert. Der Bruch sorgt dafür, dass der Wert im Definitionsbereich $[0, 1]$ des Farbwerts liegt. Der b-Wert staucht den Farbwert nach oben, so dass die Darstellung weniger rot wird. Das sollte rot-grün- oder farbblinden Menschen die interpretation erleichtern, wobei es dafür sicherlich bessere Kodierungen gibt. Das Kodierungsschema wurde ausgewählt, um die Ähnlichkeit mit dem Konzept der fMRI-Analysen zu betonen – bei der allerdings auch weitere Schemata zur Anwendung kommen.

Bemerkung 30:
Im Fall von 2-Gruppen-Vergleichen lässt sich auch die Richtung des Unterschieds in der Grafik darstellen. Bei multivariaten Testverfahren ist es nicht immer sinnvoll von einer positiven oder

negativen Richtung zu sprechen, im Fall von Korollar 8.3 kann das aber getan werden. Eine Möglichkeit im Fall von Korollar 8.3 "Richtungen" zu definieren, die auch bei Faktoren mit mehr als zwei Stufen funktionieren, ist: das Vorzeichen der Korrelation von der Variable mit ihrem Vater dafür zu nehmen – bzw. von der partiellen Korrelation bezüglich des Modells der Nullhypothese des Tests. Für die oberste getestete Ebene müsste eine willkürliche Richtung festgelegt werden. Alternativ kann auch eine feste Referenzebene vorgegeben werden, die mit Vor- und Nachfahren korreliert wird. In einer zweiten Grafik können anstatt der p-Werte auch andere Vergleichsgrößen, wie die eben erwähnten Korrelationen selbst dargestellt werden. Theoretisch könnte für ">" (bzw. "+") die obige Kodierung verwendet werden und für "<" (bzw. "-") dieselbe Kodierung mit vertauschten Rollen der Grün- und Blau-Werte. Besser wäre jedoch ein Kodierungsschema zu verwenden, dessen wahrgenommene Helligkeiten für identische p-Werte in beiden Richtungen gleich wirken.

Ringanordnung: Die Wurzel kann als Kreisscheibe im Zentrum der Grafik gezeichnet werden. Wenn die Wurzel nicht getestet wurde, beginnt die Prozedur stattdessen mit der ersten getesteten Ebene im nächsten Schritt. Direkt angrenzend kann die nächste Ebene auf einem Ring eingezeichnet werden. Die Anordnung der Taxa dieser Ebene auf dem Ring ist zunächst beliebig wählbar. Die Größe sollte die Häufigkeit der Taxa im Datensatz wiederspiegeln. Der Winkelbogen jedes Taxons sollte zunächst direkt am Ende der vorherigen beginnen. Abweichungen davon werden später angesprochen. Die nächste Ebene wird genauso aufgebaut – ihr Ring befindet sich aber weiter außen. Zur Übersichtlichkeit wird zwischen den Ringen Abstand gelassen. Die Taxa des äußeren Rings müssen so angeordnet sein, dass jede vorgestellte Linie durch den Mittelpunkt der Grafik nur durch Taxa verläuft, die eine Überkategorie des Taxons auf dem Schnitt der Linie mit dem äußersten Ring darstellt. Jede weitere Ebene wird auf dieselbe Weise als Ring hinzugefügt.

Größe der Taxa: Sei $h_j := \sum_{i=1}^{m} x_{i,j}/m$ die relative Gesamt-Relativ-Häufigkeit, mit der Taxon j in einem Versuchsdatensatz mit m Stichproben vorkommt. Hierbei werden also die Werte, die in jeder einzelnen Stichprobe die relative Häufigkeit angeben, über alle Stichproben zusammengezählt – unabhängig von ihrer Gruppenzugehörigkeit. Dann definiert sich der Winkelbogen, den j in der Grafik einnimmt durch $2\pi h_j$.

Anordnung der Taxa: Bei der Anordnung der Cluster auf einem Ring ist darauf zu achten, dass sie zu der der anderen Ringe passt. Das bedeutet, dass der Winkelbereich, den ein Kind einnimmt, in dem Winkelbereich seines Vaters liegen muss. Falls die Größe, nach der sortiert werden soll, in allen Clustern unterschiedliche Werte annimmt, kann dies gewährleistet werden, indem ein stabiler Sortier-Algorithmus in folgender Weise verwendet wird.

Definition 8.1. Stabil ist ein Sortier-Algorithmus genau dann, wenn die Reihenfolge von allen Elementen mit identischem Wert untereinander immer erhalten bleibt [Knuth (2007b)].

Eine geeignete Verwendung einer stabilen Sortierung, falls alle OTUs in jeder Ebene einem Taxon zugeordnet sind, ist:

1. Setze den äußersten Ring als aktuellen Ring.
2. Sortiere alle Cluster in ihrem Ring, die auf dem aktuellen Ring oder auf Ringen weiter außen liegen, nach dem Wert ihres Vorfahrens auf dem aktuellen Ring stabil.
3. Wenn es auf der Innenseite des aktuellen Rings keinen weiteren Ring gibt, ist der Algorithmus fertig.
4. Mache den nächst-inneren Ring des aktuellen Rings zum aktuellen Ring.
5. Gehe zu Schritt 2.

Jedes Kind liegt im Winkelbereich seines Vaters, weil die Kinder anderer Vätern, die restlichen Winkelbereiche komplett einnehmen, da die Sortierung gemäß dem Vater wegen der Reihenfolge der Sortierung immer mit der gemäß dem Kind überstimmt.

Bemerkung 31:
Sollten identische Werte in den Clustern auftreten, können diese durch Addition minimaler, für die Betrachtung irrelevanter Zahlen, unterschiedlich gemacht werden.

Das Problem gleicher Werte kann auch durch kompliziertere Vergleiche beim Sortieren sauber gelöst werden: Ein Cluster a gilt dabei als kleiner als ein Cluster b, wenn der erste (von der Wurzel betrachtet) Vorfahre A von a, der kein Vorfahre von b ist, einen kleineren Wert als der entsprechende Vorfahre B von b hat, wobei $a = A$ und $b = B$ sein darf. Falls die beiden Werte gleich sind, werden stattdessen die Namen von A und B verglichen. Die Verwendung der Namen um einen Bindungen zu brechen, ist unproblematisch, da es nur um eine Visualisierung geht. Wird diese Weise verwendet, braucht jeder Ring nur einmal sortiert zu werden.

Zur Sortierung bietet sich die relative Häufigkeit an, mit der ein Taxon insgesamt im Datensatz vorkommt.

Taxonomisch nicht vollständig zugeordnete OTUs: Bei der Darstellung der Testergebnisse ist es nicht wünschenswert, dass künstliche Cluster, die durch Kopieren von Bezeichnungen unterer Ebenen erstellt wurden, dargestellt werden. Diese sollten besser durch Lücken in den Ringen repräsentiert werden. Für andere künstliche Cluster kann die Darstellung sinnvoll sein, aber auch diese können durch Lücken wiedergegeben werden. Wenn die hierarchische Adjustierung ohne Erstellung künstlicher Cluster oder Entfernung von OTUs durchgeführt wurde, müssen die Winkelbögen der Kinder soweit verschoben (genau genommen: gedreht) werden, dass sie am Beginn des Bogens ihres Vaters starten.

Bemerkung 32:
Die grafische Darstellung kann auch ohne hierarchische Adjustierung verwendet werden, um die p-Werte explorativ darzustellen. Sie kann auch einfach nur mit gewöhnlicher Adjustierung innerhalb der Ebenen verwendet werden. Abbildung 8.1 zeigt Testresultate für Abundanzunterschiede zwischen verschiedenen Bodentypen auf diese Art. In Abbildung 8.2 sind diese Testresultate nochmal für Adjustierungen mit der hierarchischen Bedingung und mit verschiedenen Fokus-Ebenen zu sehen.

8.3. Hierarchisches Testen auf phylogenetisch transformierten Daten

Dieser Abschnitt beschreibt kurz ein Prinzip, wie eine bereits existierende phylogenetische Transformation für ein multiples Testverfahren verwendet werden kann.

Die phylogenetische Koordinatentransformationen PhILR (Abschnitt C.3.4 auf Seite 169) verbindet phylogenetische Information und Häufigkeitsdaten. Die Ergebnisse können mit multiplen Testverfahren untersucht werden. Die transformierten Variablen vermischen allerdings die ursprünglichen Variablen. Obwohl sich PhILR-transformierte Variablen biologisch interpretieren lassen, wäre auch eine direkte Interpretation im Sinne der ursprünglichen Variablen wünschenswert.

Die PhILR-Transformation überträgt die Häufigkeits-Messwerte der Blätter eines phylogenetischen Baums auf dessen innere Knoten einschließlich der Wurzel. Tests auf diesen Werten können mit dem Verfahren von Meinshausen (2008) adjustiert werden. Die Tranformation kann so definiert werden, dass nur die Messwerte aus dem Unterbaum eines Knotens für seinen Wert verwendet werden und auch keine Werte anderer Stichprobenelemente verwendet werden - abgesehen davon, dass der phylogenetische Baum davon abhängt, welche Mikroben irgendwo in der Stichprobenmenge gemessen wurden, also auf diesen bedingt werden muss. Ein univariater Test für die Hypothese der Verteilungsgleichheit verschiedener Gruppen in der transformierten Variable testet gleichzeitig auch diese Hypothese in den ursprünglichen Variablen des Unterbaums. Die Blätter erhalten keine transformierten Werte, können aber mit ihren ursprünglichen Werten in das Schema aufgenommen werden.

Die Ergebnisse wären in der Form schwerer zu visualisieren, da sich phylogenetische Bäume (als Binärbäume) nicht auf bekannte Ringe abbilden lassen. Der phylogenetische Baum kann gegebenenfalls unter Informationsverlust zu einem vorgegeben taxonomischen verkürzt werden. Knoten werden in dem Fall “verschmolzen”, indem von einer Menge von Knoten, die zu einem einzelnen Knoten verschmolzen werden soll, der gemeinsame Vorfahre in dieser Menge erhalten bleibt und alle anderen Knoten verworfen werden. Hierbei gelte jeder Knoten als einer seiner eigenen Vorfahren. Die beteiligten Kanten werden so ersetzt, dass jeder Pfad zwischen Knoten, die in beiden Bäumen vorhanden sind, sowie die Summe der beteiligten Kanten-Gewichte, erhalten bleiben.

ITS-Sequenzierungsdaten sind zur Konstruktion phylogenetischer Bäume für Pilze kaum geeignet - siehe Abschnitt C.2.1.1 auf Seite C.2.1.1. PhILR ist daher momentan eher für Bakteriendaten und zu empfehlen.

8.4. Zufällige Hypothesenmengen

In Mikrobiomanalysen hängt die Menge der Hypothesen von den Messwerten ab. Die Zusammensetzung der Hypothesenmenge könnte mit den Testergebnissen zusammenhängen, da mit sinkender Wahrscheinlichkeit des Auftretens einer Hypothese die Wahrscheinlichkeit steigen könnte, dass die Variable ungleichmäßig häufig in den zu testenden Gruppen gemessen wurde. Sicherheitshalber können Tests und Transformationen verwendet werden, die seltene Mikroben nicht zu hoch gewichten. Häufige Mikroben sollten auch mit häufigen Hypothesen assoziert sein. Das Problem zufälliger Hypothesen ist wahrscheinlich viel weniger ausgeprägt als bei den in anderen biometrischen Bereichen oft verwendeten Vorselektions-Verfahren für mehrfaktorielle Modelle, bei denen Faktoren in ein Modell aufgenommen werden, wenn sie einen niedrigen p-Wert in einem univariaten Test auf denselben Daten gezeigt haben.

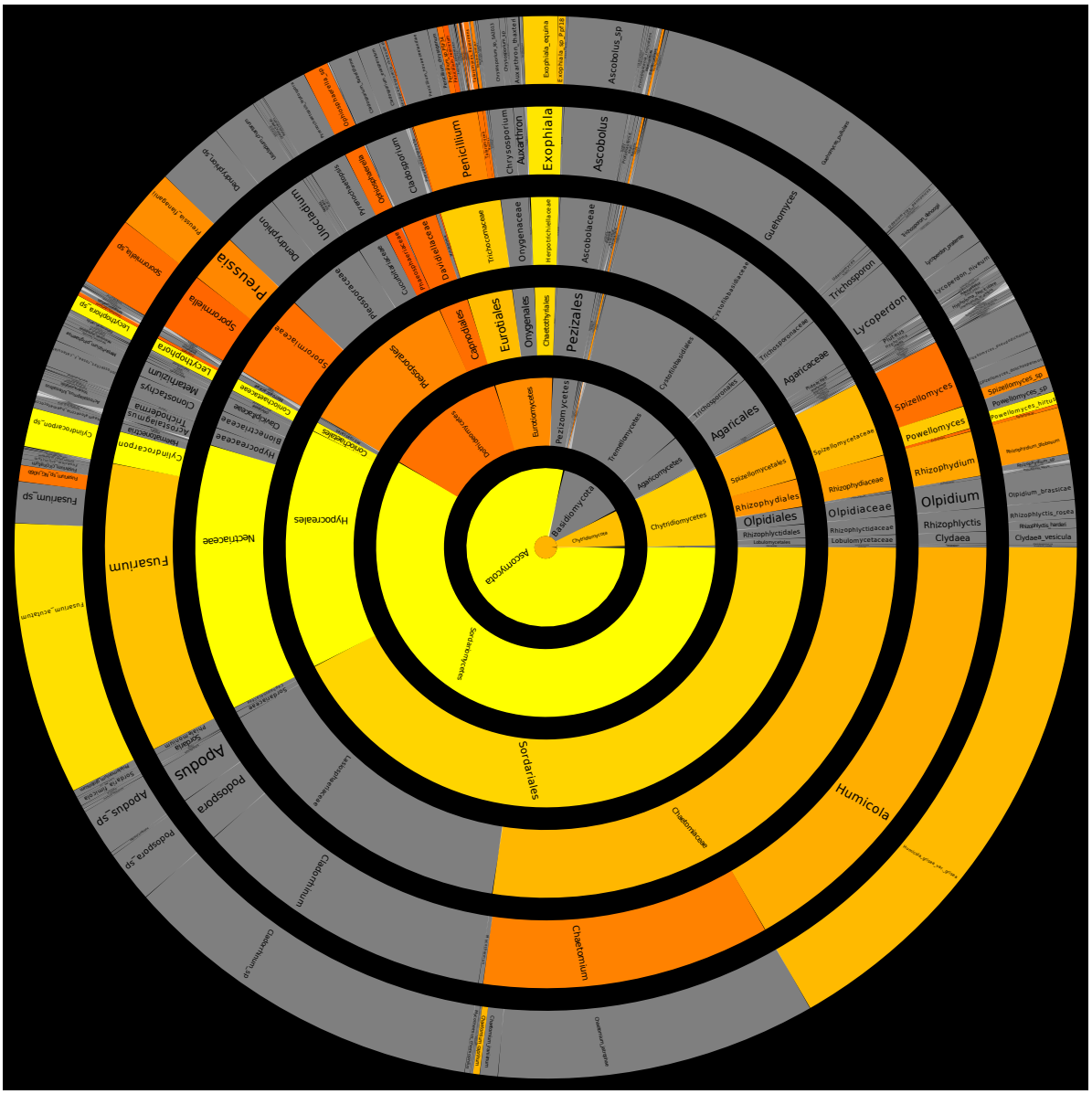

Abbildung 8.1.: Statistical Parametric Mapping übertragen auf Mikrobiomanalyseergebnisse

Nichtsignifikante Cluster sind grau. Von rot bis gelb sind p-Werte (abnehmend) dargestellt. Die Ringe entsprechen taxonomischen Ebenen. Die Wurzel des taxonomischen Baums befindet sich in der Mitte. Taxonomische Bezeichnungen der Organismengruppen sind an jeweiliger Stelle eingetragen. Auf der imaginären Verbindungslinie zwischen einer taxonomischen Gruppe und der Wurzel befinden sich die höher angesiedelten taxonomischen Zuordnungen der entsprechenden Gruppe. Die Differenz der Winkel, die eine taxonomische Gruppe begrenzen, geteilt durch einen vollständigen Kreisumfang, entspricht dem relativen Anteil dieser Gruppe im Datensatz. Der Datensatz wird in Antweiler *et al.* (2017) beschrieben. OTUs, die nicht vollständig auf allen taxonomischen Ebenen zugeordnet werden konnten, wurden vor der Analyse entfernt. Im Test wurden Häufigkeiten log-transformiert – mit Offset 0.001. Getestet wurden die Regressionskoeffizienten für den Bodenfaktor in einem Modell mit Standort-, Versuch- und Bodeneffekt. Als Test wurde der PC-Test von Läuter *et al.* (1998) mit einer Hauptkomponente verwendet. Aus der hierarchischen Adjustierung wurde hier nur die gewichtete Bonferroni-Adjustierung verwendet. Die Bedingung, dass die p-Werte von Kinder nicht kleiner sein dürfen, wurde hier nicht angewendet. Die Grafik hat an dieser Stelle explorativen Charakter.

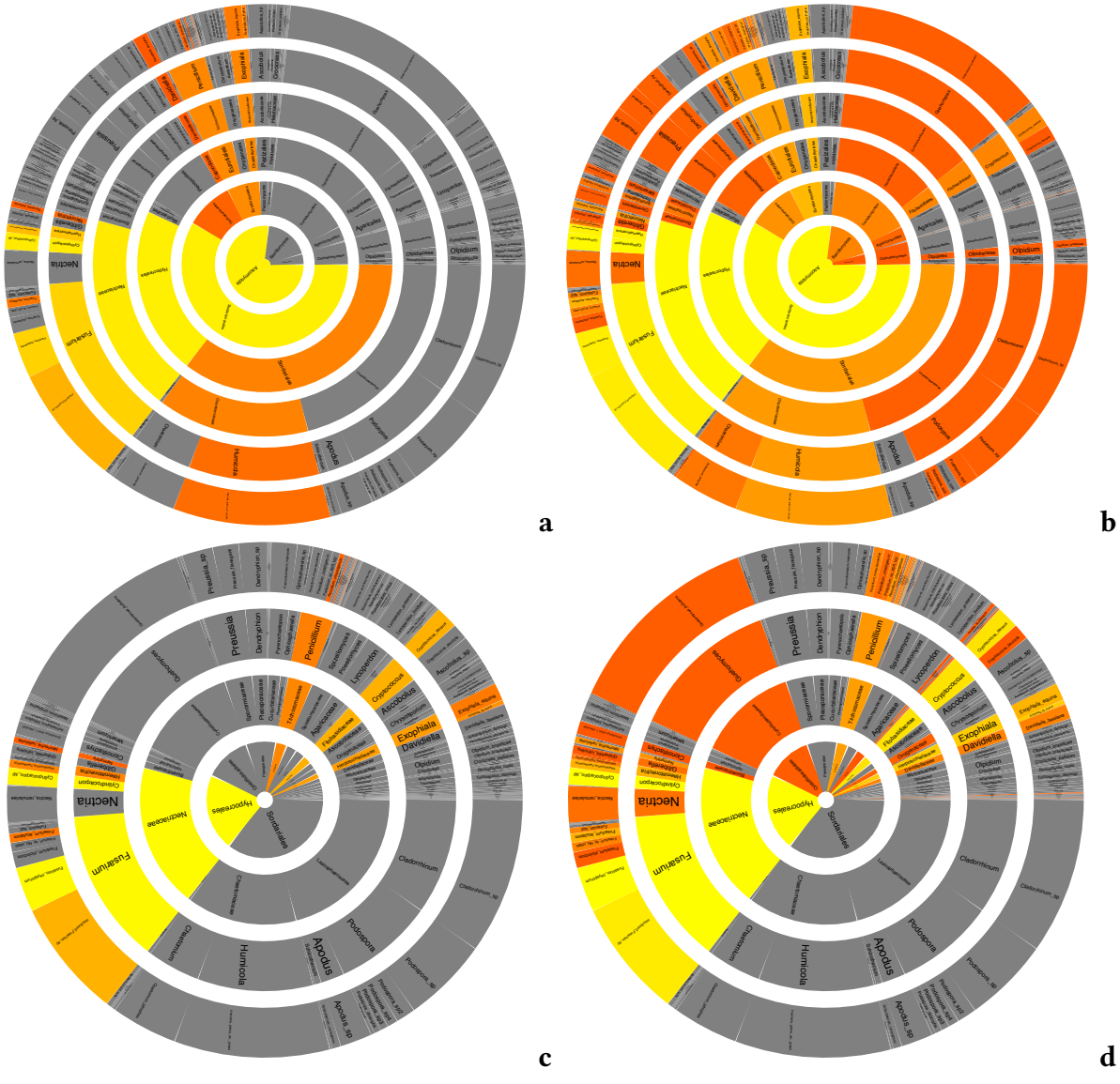

Abbildung 8.2.: Verschiedene Darstellungsoptionen

Die Legende aus Abbildung 8.1 lässt sich, bis auf wenige Änderungen, auf diese vier Grafiken übertragen. Die Daten wurden nicht log-transformiert. In **a** und **b** wurde eine hierarchische Adjustierung, basierend auf der vierten taxonomischen Ebene (Ordnung) als Fokus-Level, verwendet. In **c** und **d** wurde eine taxonomische Adjustierung gemäß Korollar 8.3 verwendet, die auf derselben Ebene beginnt, bei **d** allerdings nur teilweise - siehe unten. Bei diesen beiden Grafiken wurde dementsprechend kein PC-Test verwendet, sondern ein lineares Modell auf den Summen der relativen Häufikeiten. Die höheren Ebenen sind daher hier nicht dargestellt. Die p-Werte in **a** und **c** wurden durch ein Blatt-Anzahl-gewichtetes Bonferroni-Verfahren adjustiert und sind vollständig konfirmatorisch. Für **b** und **d** wurden diese Adjustierung ausgelassen, weshalb sie exploratisch zu verstehen sind. Die oberen Bilder unterscheiden sich in der Reihenfolge der Taxa auf ihren gemeinsamen Ringen von den unteren Bildern, da in den oberen Bildern die höchsten drei taxonomischen Ebenen bei der Sortierung berücksichtigt wurden, die nicht Teil der unteren Bilder sind.

Teil IV.

Diskussion

9. Diskussion

In der vorliegenden Arbeit wurden erfolgreich mehrere Auswertemethoden für Mikrobiomdaten entworfen, die biologische Konzepte berücksichtigen. Besonderer Wert wird auf die Interpretierbarkeit der Ergebnisse gelegt. Für Äquivalenztests wurde die Verwendung etablierter ökologischer Abstandsmaße vorgeschlagen. Hierzu wurden Verfahren kombiniert und verbessert, damit unter realistischen Stichprobengrößen und Versuchsbedingungen getestet werden kann. Das betrifft insbesondere die Varianzbestimmung durch ein stratifiziertes Jackknife-Verfahren zur konservativen Varianzbestimmung, das sich auch für kleine Stichproben eignet. Die zuvor veröffentlichte Form schränkte das Verhältnis der ausgelassenen Teile in dem Varianzschätzer zwischen Straten soweit ein, dass die Gruppengrößenrelationen nur wenige sehr genaue Werte annehmen durften. Selbst durch eine gute Versuchsplanung kann dies nicht garantiert werden, wie sich auch in dem Versuch gezeigt hat, der im Rahmen der Projektförderung der vorliegenden Dissertation durchgeführt wurde. Der Nachweis, dass das hier angepasste Verfahren die gewünschten Eigenschaften besitzt, wurde mithilfe einer ANOVA-Typ-Zerlegung des Varianzschätzers durchgeführt. Für die Ausmaße der Konservativität des Varianzschätzers wurden, für den Spezialfall einer linearen Statistik, Gleichungen aufgestellt. Der Varianzschätzer wurde dann zur Konstruktion eines Konfidenzintervalls und dieses wiederum als Grundlage für einen Äquivalenztest verwendet. Wie nötige Information über Äquivalenzgrenzen in verschiedenen Versuchsschemen gewonnen werden kann, wurde beschrieben. Für Tests auf Unterschied zwischen Mikrobenhäufigkeiten wurde bei zwei multiplen Testverfahren phylogenetische Information in die Analysen integriert. In hierarchischen Tests wurde die taxonomische Struktur als Hierarchie verwendet und die Hypothesenformulierung angepasst. Auf diese Weise wurde insbesondere die Auswertung auf kommunizierbare Organismengruppen konzentriert. Zu den Ergebnissen wurde eine übersichtliche, grafische Darstellungsmöglichkeit entwickelt. Für ein sequenzielles Testverfahren, wurden Gütevorteile beschrieben, die nicht auf der Korrelationsstruktur der Variablen beruhen und durch phylogenetische Information möglich sind.

Güte-Nachweis-Problematik: Einen Güte-Vorteil durch die Ausnutzung phylogenetischer Information bei den multiplen Testverfahren nachzuweisen, ist problematisch, da unklar ist, auf welche Mikroben ein Versuch einen Einfluss haben sollte und auf welche nicht. Der Wahrheitsgehalt der Hypothesen ist unklar. Auf der anderen Seite ist dieser zwar bei Simulationen bekannt, aber hier muss ein Zusammenhang zwischen den Häufigkeiten und der Phylogenie vorgegeben werden. Wie dieser realistisch auszusehen hat, ist schwer abzusehen. Daher wurden hier bekannte Zusammenhänge möglichst umfassend zusammengetragen und berücksichtigt. Außerdem wurde mehr Wert auf die Interpretierbarkeit der Ergebnisse gelegt.

Auf diese Weise konnten relevante Fortschritte erzielt werden.

Eine Ausnahme der Nachweisproblematik entsteht durch einen anderen Fokus: Ein Güte-Vorteil kann durch Annäherungen realer Verhältnisse an die Bedingungen eines Verfahrens entstehen. Hier trifft das insbesondere für das sequenzielle Testen mit datenabhängig geordneten Hypothesen nach Kropf zu. Ein Zusammenhang zwischen phylogenetischer Information und den Mikrobenhäufigkeiten braucht nicht angenommen zu werden. Der Wahrheitsgehalt der Hypothesen braucht daher nicht bekannt zu sein. Wenn die Varianzen in verwandten Gruppen von Mikroben ähnlicher sind als in gleich-großen anderen Gruppen, kann das Verfahren davon profitieren, solche Gruppen nach Verwandtschaft zu bilden.

Testgültigkeit: Bei einstelligen Stichprobengrößen und überwiegend 0-Einträgen in den Realisierungen der Zufallsvektoren stellt sich die Frage, ob sich Tests asymptotisch rechtfertigen lassen. Nichtasymptotische Randomisierungsargumente, mit denen Permutationstests begründet werden können, sind ebenfalls problematisch. Sie basieren auf Gruppenzugehörigkeiten: Die Hypothese besagt, dass die Zuordnung der Gruppe zu den Werten zufällig war. Diese Hypothese kann nicht für eine Dimension des Vektors wahr und für eine andere falsch sein. Da die Verteilungen einer Untermenge von Variablen sich nur zwischen Gruppen unterscheiden können, wenn die Randomisierungshypothese falsch ist, kann das Testverfahren auch als Test für die Hypothese der Gleichheit der Verteilungen dieser Untermenge interpretiert werden. Für diesen sind einige Adjustierungsverfahren wie z.B. Bonferroni anwendbar.

Für multiple Testschemata auf ungewöhnlichen Daten bei kleinen Stichprobengrößen ist eine begründete Testauswahl unüblich. Hier wurde ein lineares Modell (bzw. PC-Test [Läuter *et al.* (1998)] bei mehreren Dimensionen) verwendet, weil ich dieses standardmäßig oft verwende. Die Score-Bildung ist sicher günstig für das asymptotische Argument, allerdings stellt sich die Frage, ob das Bilden des Scores gerechtfertigt ist. Die Frage lässt sich auch für die Variablensortierung in dem sequenziellen Testverfahren stellen. Andererseits basiert der in der Ökologie weit-verbreitete Bray-Curtis Abstand ebenfalls auf einer Linearkombination.

Zusatznutzen phylogenetischer Information: Für statistische Verfahren auf hochdimensionalen Daten ist meist zu erwarten, dass sie davon profitieren, wenn weniger Skalierungsunterschiede zwischen den Variablen vorliegen, die nicht auf den Versuchseinflüssen beruhen. Bereits aus messtechnischer Sicht ist momentan bei den meisten Mikrobiomanalysen zu erwarten, dass die Variablen unterschiedlich skaliert sein werden, denn: 1. unterscheidet sich die Anzahl der Markergene, die sequenziert werden können, zwischen den Organismen und 2. werden die DNA-Fragmente, abhängig von ihrer Sequenz, unterschiedlich leicht aufgegriffen. Beides sind Effekte, die zwischen näher-verwandten Organismen weniger zu erwarten sind. Die Hauptkomponenten-Analysen der vorliegenden Daten stimmten besser mit dem Versuchsdesign überein, wenn sie auf Kovarianzmatrizen als auf Korrelationsmatrizen aufbauten. Eine naive Skalenangleichung scheint hier also kontraproduktiv zu sein.

Es liegt somit eine Situation mit Potential für phylogenetische Skalenangleichungs-Ansätze vor.

Unabhängig davon kann ein Zusammenhang zwischen Verwandtschaft und Auswirkung einer Behandlung vermutet werden. In ökologischen Studien ist die Verwendung phylogenetischer Abstände üblich.

Ein Fokus auf Testergebnisse, die gängigen Einteilungen entsprechen, erzeugt viele Vorteile. Die Ergebnisse lassen sich besser kommunizieren und verstehen. Eine sinnvolle Einteilung kann ihre Sinnhaftigkeit auf die Ergebnisse übertragen. Ein positives Testergebnis für eine verwandte Gruppe von Organismen kann in einem Folgeversuch (ggf. Auftauung von Proben) genauer untersucht werden, bei dem Primer oder Marker oder Kultivierungsverfahren speziell für diese ausgewählt werden.

Mikrobiomanalysen sind ökologische Untersuchungen von mikrobiellen Gemeinschaften. Dennoch werden viele Ansätze verfolgt, die die Häufigkeiten einzelner Mikroben als abstrakte Variablen behandeln, denen nur lineare Abhängigkeiten zugrunde liegen, die auch erst aus ihren Werten abgelesen werden können. Der Anteil solcher Ansätze scheint höher zu sein als der, Fotos ohne Beachtung räumlicher Zusammenhänge zu analysieren. (Wie viele blaue, grüne, gelbe, usw. Pixel gibt es in dem ersten Foto und wie viele in dem zweiten?) Das Ausnutzen phylogenetischer Information wird möglicherweise nie die Bedeutung haben, die räumliche Nähe in Fotos zukommt. Dies ist eher für ökologische Zusammenhänge zu erwarten. Dennoch scheint es mir ein Schritt in die richtige Richtung zu sein, der auch Probleme offen legt und das Verständnis über die Struktur der Proben verbessern kann. Bei metagenomischen Analysen fallen viele Probleme weg, da die phylogenetische Verwandtschaften einzeln für jedes Gen betrachtet werden können.

Informationstheoretische Ansätze: Ein Teil der Forschungsarbeit verfolgte ursprünglich Ansätze aus dem Themengebiet der Informationstheorie. Die Mikrobengemeinschaften wurden hierbei als empirische Verteilungen angesehen. Da bei den Ansätzen nicht der gewünschte Nutzen zu erkennen war und eine umfangreichere Beschreibung teilweise fortgeschrittener, informationstheoretischer Grundlagen nötig gewesen wäre, wurden sie nicht mit in den vorliegenden Text übernommen.

Ungünstige Daten: Mit Beginn des Dissertationsvorhabens wurde der Versuch geplant, aus dem schließlich die ersten Daten kamen. Bis diese vorlagen, hat es einige Jahre gedauert. Diese Daten stammten von Pilzen und waren lange Zeit die einzigen, die verwendet wurden. Phylogenetische Bäume lagen nicht vor. Für einige Verfahren wurde auf taxonomische Bäume ausgewichen, die aus Datenbanken bezogen wurden. Dementsprechend stand nicht das volle Potential der theoretisch möglichen Information zur Verfügung. Auch die Anzahl der Variablen verringerte sich auf die, die in Datenbanken klassifiziert waren. Die meiste Zeit wurde versucht, Methoden anhand einer Sequenz-Ähnlichkeitsmatrix von Pilzen zu entwerfen. Das war oft nicht erfolgreich. Nachdem ich von den speziellen Problemen bei Pilzen

erfahren habe, scheiterte der Versuch, ein Hauptkomponentenverfahren anhand von Bakteriendaten zu bestätigen. Ein großer Teil der vorliegenden Arbeit besteht in der Erklärung der Datenstruktur. Leider fehlte oft die nötige Zeit, Verfahren an diese anzupassen, weshalb dies auf weitere aufbauende Arbeiten verschoben werden muss.

Auf die Beschreibung der Daten und ihre Gewinnung wurde daher (und wegen der sich abzeichnenden langsamen Entwicklung hin zu Metagenom-Sequenzierung) viel Wert gelegt. Die elementare Struktur der Sequenzen und Fehlerquellen durch Sequenzierungen wurden besprochen. Die komplexe Verwandtschaftsbeziehung zwischen Variablen wurde näher betrachtet. Darauf aufbauend hätte idealerweise eine Suche nach bestmöglichen Datensätzen stattfinden sollen.

Aussicht: In den letzten Jahren konnte die Anzahl der 0-Werte und Variablen durch geschicktere Betrachtung der Daten deutlich reduziert werden. Durch Fortschritte in der Messtechnik wurden die Messungen günstiger und länger. Mit den Längen nimmt auch ihre phylogenetische Aussagekraft zu. Es bleibt zu hoffen, dass zukünftig mittelfristig Markergene gewählt werden, die nur in einer Kopie vorkommen, bis das Sequenzieren ganzer Genome dominieren wird – was allerdings einen der Vorteile der hier entwickelten Ansätze reduziert. In Bereichen, in denen der Versuchsaufbau nicht die Kosten dominiert, können Stichprobenumfänge größer werden. Proben können mehrfach gemessen und Referenzproben verwendet werden. Die Messung von Zeitreihen ermöglicht andere Versuchsdesigns. Insbesondere kann sich Information über ökologische Dynamiken ansammeln. Es könnten Laborversuche ohne Störgrößen geplant werden, die Dynamiken untersuchen und ökologische Information schaffen, die in allen weiteren Auswertungen berücksichtigt werden kann. Anteilsmäßig wird die Finanzierung wahrscheinlich eher größere medizinische Projekte betreffen, in denen beispielsweise Patienten ihre Stuhlproben selbst vorbeibringen – mit unterschiedlichen Zeitlängen und Bedingungen, bis eine oftmals unzureichende Kühlung beginnt, die nicht den Qualitätsstandards biologischer Studien genügt, womit selbst die Wiederauswertung mit zukünftig verbesserten Verfahren für diese Datenerhebungen in Frage gestellt wird.

Anwendungsfelder: Die Methodiken von Mikrobiomanalysen lassen sich auf größere mehrzellige Organismen verallgemeinern. Das wurde hier nicht ausgeführt und hätte den Text deutlich verlängert.

Pflanzenpollen können prinzipiell mit denselben Methoden untersucht werden. Durch die Analyse von beispielsweise Bienenwaben können Aussagen über Zusammensetzung und Veränderungen in großflächigen Landabschnitten und Zusammenhänge zwischen den Bestäubern und Pflanzen getroffen werden [Pornon *et al.* (2017)].

Auch Mitochondrien lassen sich mit den Methoden der Mikrobiomanalyse untersuchen. Eine eukaryotische Zelle enthält eine ganze Population von Mitochondrien. In menschlichen Zellen sammeln sich Mutation in Mitochondrien vor allem in Gehirn und Herzmuskel älterer Individuen an [Burger *et al.* (2003)]. Das deutet auf Forschungspotential hin, wenn günstig lange Sequenzen bestimmt werden können.

Zellen des Immunsystems sind relativ leicht zugänglich. Next-Generation Ansätze könnten eventuell akute Infektabwehr und latente Immunität auseinanderhalten und identifizieren – ohne dass man sich vorher auf bestimmte Erreger festlegen müsste.

Momentan schon praktikabel, aber noch zu teuer für Routine-Untersuchungen, ist der schnelle sequenzbasierte Nachweis von beliebigen Erregern in einer Probe, in der Erreger vermutet werden. Für seltene Infektionskrankheiten wurde bereits nachgewiesen, dass Next Generation Sequencing Verfahren traditionellen Diagnosemethoden überlegen sind [Tiew *et al.* (2020)].

Next Generation Sequencing basierte Ansätze zeigen daher und aus vielen weiteren Gründen ein weites Anwendungsspektrum, dessen Bedeutung weiter anwachsen wird. Diese Arbeit leistet einen kleinen Betrag für darauf basierende Datenauswertungen.

Literaturverzeichnis

Amir, Amnon, McDonald, Daniel, Navas-Molina, Jose A, Kopylova, Evguenia, Morton, James T, Xu, Zhenjiang Zech, Kightley, Eric P, Thompson, Luke R, Hyde, Embriette R, Gonzalez, Antonio, *et al.* 2017. Deblur rapidly resolves single-nucleotide community sequence patterns. *MSystems*, **2**(2), e00191–16.

Antweiler, Kai, Schreiter, Susanne, Keilwagen, Jens, Baldrian, Petr, Kropf, Siegfried, Smalla, Kornelia, Grosch, Rita, *&* Heuer, Holger. 2017. Statistical test for tolerability of effects of an antifungal biocontrol strain on fungal communities in three arable soils. *Microbial biotechnology*, **1**0(2), 434–449.

Arvesen, J.N. 1969. Jackknifing U-statistics. *Ann. Math. Statist.*, 2076–2100.

Begon, Michael, Howarth, Robert W, *&* Townsend, Colin R. 2016. *Ökologie*. Third edn. Springer-Verlag.

Bennett, Simon. 2004. Solexa ltd. *Pharmacogenomics*, **5**(4), 433–438.

Berger, Roger L. 1982. Multiparameter hypothesis testing and acceptance sampling. *Technometrics*, **2**4(4), 295–300.

Betts, Holly C., Puttick, Mark N., Clark, James W., Williams, Tom A., Donoghue, Philip C. J., *&* Pisani, Davide. 2018. Integrated genomic and fossil evidence illuminates life's early evolution and eukaryote origin. *Nature Nature Ecology & Evolution*, 1556–1562.

Bolyen, Evan, Rideout, Jai Ram, Dillon, Matthew R, Bokulich, Nicholas A, Abnet, Christian C, Al-Ghalith, Gabriel A, Alexander, Harriet, Alm, Eric J, Arumugam, Manimozhiyan, Asnicar, Francesco, *et al.* 2019. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nature biotechnology*, **3**7(8), 852–857.

Bonfante, Paola, *&* Anca, Iulia-Andra. 2009. Plants, mycorrhizal fungi, and bacteria: a network of interactions. *Annual review of microbiology*, **6**3, 363–383.

Bullerwell, Charles E, *&* Lang, B Franz. 2005. Fungal evolution: the case of the vanishing mitochondrion. *Current opinion in microbiology*, **8**(4), 362–369.

Burger, Gertraud, Gray, Michael W, *&* Lang, B Franz. 2003. Mitochondrial genomes: anything goes. *Trends in genetics*, **1**9(12), 709–716.

Cadotte, Marc W, Jonathan Davies, T, Regetz, James, Kembel, Steven W, Cleland, Elsa, *&* Oakley, Todd H. 2010. Phylogenetic diversity metrics for ecological communities: integrating species richness, abundance and evolutionary history. *Ecology letters*, **1**3(1), 96–105.

Callahan, Benjamin J, McMurdie, Paul J, Rosen, Michael J, Han, Andrew W, Johnson, Amy Jo A, & Holmes, Susan P. 2016. DADA2: high-resolution sample inference from Illumina amplicon data. *Nature methods*, **1**3(7), 581.

Campbell, Neil A., Reece, Jane B., & Mitchell, Lawrence G. 1999. *Biology*. Fifth edn. Addison-Wesley.

Caporaso, J Gregory, Kuczynski, Justin, Stombaugh, Jesse, Bittinger, Kyle, Bushman, Frederic D, Costello, Elizabeth K, Fierer, Noah, Pena, Antonio Gonzalez, Goodrich, Julia K, Gordon, Jeffrey I, *et al.* 2010. QIIME allows analysis of high-throughput community sequencing data. *Nature methods*, **7**(5), 335.

Case, Rebecca J, Boucher, Yan, Dahllöf, Ingela, Holmström, Carola, Doolittle, W Ford, & Kjelleberg, Staffan. 2007. Use of 16S rRNA and rpoB genes as molecular markers for microbial ecology studies. *Applied and environmental microbiology*, **7**3(1), 278–288.

Chen, Zhao-Jin, Xu, Ge, Ding, Chuan-Yu, Zheng, Bao-Hai, Chen, Yan, Han, Hui, Li, Yu-Ying, Shi, Jian-Wei, & Hu, Lan-Qun. 2020. Illumina MiSeq sequencing and network analysis the distribution and co-occurrence of bacterioplankton in Danjiangkou Reservoir, China. *Archives of Microbiology*, 1–15.

Cormen, Thomas H., Leiserson, Charles E., Rivest, Ronald L., & Stein, Clifford. 2001. *Introduction to Algorithms*. Second edn. MIT-Press.

De Caceres, Miquel, Legendre, Pierre, & He, Fangliang. 2013. Dissimilarity measurements and the size structure of ecological communities. *Methods in ecology and evolution*, **4**(12), 1167–1177.

Dutheil, Julien Y (ed). 2020. *Statistical Population Genomics*. Springer.

Edgington, Eugene. 1995. *Randomization tests*. Third edn. Marcel Dekker, inc.

Efron, B., & Stein, C. 1981. The jackknife estimate of variance. *Ann. Math. Statist.*, 586–596.

Efron, B., & Tibshirani, R. (eds). 1993. *An Introduction to the Bootstrap*. Chapman and Hall/CRC.

Eisenstein, Michael. 2012. *Oxford Nanopore announcement sets sequencing sector abuzz*.

Faith, Daniel P. 1992. Conservation evaluation and phylogenetic diversity. *Biological conservation*, **6**1(1), 1–10.

Federhen, Scott. 2011. The NCBI taxonomy database. *Nucleic acids research*, **4**0(D1), D136–D143.

Goeman, Jelle J, & Mansmann, Ulrich. 2008. Multiple testing on the directed acyclic graph of gene ontology. *Bioinformatics*, 537–455.

Graw, Jochen. 2015. *Genetik*. 6 edn. Springer.

Hajek, J. 1968. Asymptotic normality of simple linear rank statistics under alternatives. *Ann. Math. Statist.*, 325–346.

Hastings, P. J., Lupski, James R., Rosenberg, Susan M., & Ira, Grzegorz. 2009. Mechanisms of change in gene copy number. *Nat Rev Genet*, **1**0(8), 551–564.

Heidelberg, John F, Eisen, Jonathan A, Nelson, William C, Clayton, Rebecca A, Gwinn, Michelle L, Dodson, Robert J, Haft, Daniel H, Hickey, Erin K, Peterson, Jeremy D, Umayam, Lowell, *et al.* 2000. DNA sequence of both chromosomes of the cholera pathogen Vibrio cholerae. *Nature*, **4**06(6795), 477.

Hein, Jotun, Schierup, Mikkel H., & Carsten, Wiuf. 2005. *Gene Genealogies, Variation and Evolution.* Oxford University Press.

Hirsch-Kauffmann, Monica, & Schweiger, Manfred. 1992. *Biologie für Mediziner und Naturwissenschaftler.* Second edn. Thieme.

Hoeffding, W. 1948. A class of statistics with asymptotically normal distributions. *Ann. Math. Statist.*, 293–325.

Jay, Zackary J, & Inskeep, William P. 2015. The distribution, diversity, and importance of 16S rRNA gene introns in the order Thermoproteales. *Biology direct*, **1**0(1), 35.

Karlin, Samuel, & Rinott, Yosef. 1982. Applications of ANOVA Type Decompositions for Comparisons of Conditional Variance Statistics including Jackknife Estimates. *Ann. Statist. 10*, 485–501.

Kaufmann, H., & A., Hädener. 1996. *Grundlagen der organischen Chemie.* 10th edn. Birkhäuser Verlag.

King, Jack Lester, & Jukes, Thomas H. 1969. Non-darwinian evolution. *Science*, **1**64(3881), 788–798.

Knight, Rob, Vrbanac, Alison, Taylor, Bryn C, Aksenov, Alexander, Callewaert, Chris, Debelius, Justine, Gonzalez, Antonio, Kosciolek, Tomasz, McCall, Laura-Isobel, McDonald, Daniel, *et al.* 2018. Best practices for analysing microbiomes. *Nature Reviews Microbiology*, 1.

Knippers, Rolf. 1997. *Molekulare Genetik.* Seventh edn. Thieme.

Knuth, Donald. 2007a. *The Art of Computer Programming Volume 1.* Third edn. Addison-Wesley.

Knuth, Donald. 2007b. *The Art of Computer Programming Volume 3.* Second edn. Addison-Wesley.

Kropf, Siegfried. 2000. *Hochdimensionale multivariate Verfahren in der medizinischen Statistik.* Shaker.

Labrie, Simon J, Samson, Julie E, & Moineau, Sylvain. 2010. Bacteriophage resistance mechanisms. *Nature Reviews Microbiology*, **8**(5), 317.

Lewin, Benjamin. 2000. *Genes*. Seventh edn. Oxford University Press.

Lloyd, Karen G, Ladau, Josh, Steen, Andrew D, Yin, Junqi, & Crosby, Lonnie. 2018. Phylogenetically novel uncultured microbial cells dominate Earth microbiomes. *bioRxiv*, 303602.

Lozupone, Catherine, & Knight, Rob. 2005. UniFrac: a new phylogenetic method for comparing microbial communities. *Appl. Environ. Microbiol.*, **7**1(12), 8228–8235.

Lozupone, Catherine A, Hamady, Micah, Kelley, Scott T, & Knight, Rob. 2007. Quantitative and qualitative β diversity measures lead to different insights into factors that structure microbial communities. *Appl. Environ. Microbiol.*, **7**3(5), 1576–1585.

Lozupone, Catherine A, Stombaugh, Jesse I, Gordon, Jeffrey I, Jansson, Janet K, & Knight, Rob. 2012. Diversity, stability and resilience of the human gut microbiota. *Nature*, **4**89(7415), 220.

Lucius, Richard, Loos-Frank, Brigitte, & Lane, Richard P. 2018. *Biologie von Parasiten*. Third edn. Springer.

Läuter, Jürgen, Glimm, Ekkehard, & Kropf, Siegfried. 1998. Multivariate tests based on left-spherically distributed linear scores. *Annals of Statistics*, 1972–1988.

Margulies, Marcel, Egholm, Michael, Altman, William E, Attiya, Said, Bader, Joel S, Bemben, Lisa A, Berka, Jan, Braverman, Michael S, Chen, Yi-Ju, Chen, Zhoutao, *et al.* 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **4**37(7057), 376.

Martiny, JBH, Jones, SE, Lennon, JT, & Martiny, AC. 2015. *Microbiomes in light of traits: a phylogenetic perspective. Science 350: aac9323.*

McCarthy, Alice. 2010. Third generation DNA sequencing: pacific biosciences' single molecule real time technology. *Chemistry & biology*, **1**7(7), 675–676.

Medini, Duccio, Donati, Claudio, Tettelin, Herve, Masignani, Vega, & Rappuoli, Rino. 2005. The microbial pan-genome. *Current opinion in genetics & development*, **1**5(6), 589–594.

Meinshausen, Nicolai. 2008. Hierarchical testing of variable importance. *Biometrika*, **9**5(2), 265–278.

Metzker, Michael L. 2010. Sequencing technologies—the next generation. *Nature reviews genetics*, **1**1(1), 31.

Nagy, László G., Kocsubé, Sándor, Csanádi, Zoltán, Kovács, Gábor M., Petkovits, Tamás, Vágvölgyi, Csaba, & Papp, Tamás. 2012. Re-Mind the Gap! Insertion – Deletion Data Reveal Neglected Phylogenetic Potential of the Nuclear Ribosomal Internal Transcribed Spacer (ITS) of Fungi. *PLOS ONE*, **7**(11), 1–9.

Neilson, Julia W, Califf, Katy, Cardona, Cesar, Copeland, Audrey, van Treuren, Will, Josephson, Karen L, Knight, Rob, Gilbert, Jack A, Quade, Jay, Caporaso, J Gregory, *et al.* 2017. Significant Impacts of Increasing Aridity on the Arid Soil Microbiome. *MSystems*, **2**(3), e00195–16.

Nentwig, Wolfgang, Bacher, Sven, & Brandl, Roland. 2011. *Ökologie kompakt*. Third edn. Springer.

Paetzold, Bernhard, Willis, Jesse R, de Lima, Joao Pereira, Knölseder, Nastassia, Brüggemann, Holger, Quist, Sven R, Gabaldon, Toni, & Güell, Marc. 2019. Skin microbiome modulation induced by probiotic solutions. *Microbiome*, **7**(1), 95.

Parks, Donovan H, Chuvochina, Maria, Waite, David W, Rinke, Christian, Skarshewski, Adam, Chaumeil, Pierre-Alain, & Hugenholtz, Philip. 2018. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nature biotechnology*.

Philippot, Laurent, Raaijmakers, Jos M, Lemanceau, Philippe, & Van Der Putten, Wim H. 2013. Going back to the roots: the microbial ecology of the rhizosphere. *Nature Reviews Microbiology*, **1**1(11), 789.

Poret-Peterson, Amisha T, Sayed, Nada, Glyzewski, Nathaniel, Forbes, Holly, Gonzalez-Orta, Enid T, & Kluepfel, Daniel A. 2019. Temporal Responses of Microbial Communities to Anaerobic Soil Disinfestation. *Microbial Ecology*, 1–11.

Pornon, Andre, Andalo, Christophe, Burrus, Monique, & Escaravage, Nathalie. 2017. DNA metabarcoding data unveils invisible pollination networks. *Scientific reports*, **7**(1), 16828.

Qiu, Zhiying. 2014. *FWER CONTROLLING PROCEDURES FOR TESTING MULTIPLE HYPOTHESES WITH HIERARCHICAL STRUCTURE AND APPLICATIONS IN CLINICAL TRIALS*. Ph.D. thesis, New Jersey Institute of Technology.

Quenouille, M.H. 1949. Approximate tests of correlation in time series. *J.R.Stat.Soc.B.*, 68–84.

Quince, Christopher, Walker, Alan W, Simpson, Jared T, Loman, Nicholas J, & Segata, Nicola. 2017. Shotgun metagenomics, from sampling to analysis. *Nature biotechnology*, **3**5(9), 833.

Ramisetty, Bhaskar Chandra Mohan, & Sudhakari, Pavithra Anantharaman. 2019. Bacterial 'Grounded' Prophages: Hotspots for Genetic Renovation and Innovation. *Frontiers in Genetics*, **1**0, 65.

Raven, Peter, Evert, Ray F., & Eichhorn, Susan E. 2000. *Biology of Plants*. Sixth edn. W.H. Freeman and Company Worth Publishers.

Rokas, Antonis, Wisecaver, Jennifer H, & Lind, Abigail L. 2018. The birth, evolution and death of metabolic gene clusters in fungi. *Nature Reviews Microbiology*, 1.

Sanjuan, Rafael, & Domingo-Calap, Pilar. 2016. Mechanisms of viral mutation. *Cell Mol Life Sci*, **7**3(23), 4433–4448. 27392606[pmid].

Sauermost, Rolf, & Freudig, Doris. 1999. *Lexikon der Biologie.* Spektrum, Akad. Verlag.

Schoch, Conrad L., Seifert, Keith A., Huhndorf, Sabine, Robert, Vincent, Sponge, John L., Levesque, C. Andre, & Chen, Wen. 2012. Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proceedings of the National Academy of Sciences*, **1**09(16), 6241–6246.

Scholz, F. W. 2007. *Technical Report: The bootstrap small sample properties.* Tech. rept. University of Washington.

Schreiter, Susanne, Ding, Guo-Chun, Grosch, Rita, Kropf, Siegfried, Antweiler, Kai, & Smalla, Kornelia. 2014. Soil type-dependent effects of a potential biocontrol inoculant on indigenous bacterial communities in the rhizosphere of field-grown lettuce. *FEMS microbiology ecology*, **9**0(3), 718–730.

Schulz, G. E., & Schrimer, R. H. 1979. *Principles of Protein Structure.* Springer.

Shaffer, J. P. 1995. Multiple Hypothesis Testing. *Ann. Rev. Psych.*, 561–584.

Shanahan, Fergus. 2002. The host–microbe interface within the gut. *Best practice & research Clinical gastroenterology*, **1**6(6), 915–931.

Shannon, Claude Elwood. 1948. A mathematical theory of communication. *Bell system technical journal*, **2**7(3), 379–423.

Shao, Jun. 2003. *Mathematical Statistics.* Second edn. Springer.

Shendure, Jay, & Ji, Hanlee. 2008. Next-generation DNA sequencing. *Nature biotechnology*, **2**6(10), 1135.

Silverman, Justin D, Washburne, Alex D, Mukherjee, Sayan, & David, Lawrence A. 2017. A phylogenetic transform enhances analysis of compositional microbiota data. *Elife*, **6**, e21887.

Sinha, Rahul, Stanley, Geoff, Gulati, Gunsagar Singh, Ezran, Camille, Travaglini, Kyle Joseph, Wei, Eric, Chan, Charles Kwok Fai, Nabhan, Ahmad N, Su, Tianying, Morganti, Rachel Marie, *et al.* 2017. Index switching causes "spreading-of-signal" among multiplexed samples in Illumina HiSeq 4000 DNA sequencing. *BioRxiv*, 125724.

Storch, Volker, Welsch, Ulrich, & Wink, Michael. 2013. *Evolutionsbiologie.* Third edn. Springer-Verlag.

Tiew, Pei Yee, Mac Aogain, Micheal, Ali, Nur Atikah Binte Mohamed, Thng, Kai Xian, Goh, Karlyn, Lau, Kenny JX, & Chotirmall, Sanjay H. 2020. The Mycobiome in Health and Disease: Emerging Concepts, Methodologies and Challenges. *Mycopathologia*, 1–25.

Tukey, J. 1958. Bias and confidence in not-quite large samples (Abstract). *Ann. Math. Statist.*, 614.

Venail, Patrick A, Narwani, Anita, Fritschie, Keith, Alexandrou, Markos A, Oakley, Todd H, & Cardinale, Bradley J. 2014. The influence of phylogenetic relatedness on species interactions among freshwater green algae in a mesocosm experiment. *Journal of Ecology*, **1**02(5), 1288–1299.

Wagner, Patrick L, & Waldor, Matthew K. 2002. Bacteriophage control of bacterial virulence. *Infection and immunity*, **7**0(8), 3985.

Walker, Alan W, Martin, Jennifer C, Scott, Paul, Parkhill, Julian, Flint, Harry J, & Scott, Karen P. 2015. 16S rRNA gene-based profiling of the human infant gut microbiota is strongly influenced by sample processing and PCR primer choice. *Microbiome*, **3**(1), 26.

Wehner, Rüdiger, & Gehring, Walter. 1993. *Zoologie*. 23rd edn. Thieme.

Wilmes, Paul, Simmons, Sheri L, Denef, Vincent J, & Banfield, Jillian F. 2008. The dynamic genetic repertoire of microbial communities. *FEMS microbiology reviews*, **3**3(1), 109–132.

Wisecaver, Jennifer H., & Rokas, Antonis. 2015. Fungal metabolic gene clusters—caravans traveling across genomes and environments. *Frontiers in Microbiology*, **6**, 161.

Wu, Dongying, Doroud, Ladan, & Eisen, Jonathan A. 2013. TreeOTU: operational taxonomic unit classification based on phylogenetic trees. *arXiv preprint arXiv:1308.6333*.

Yarza, Pablo, Yilmaz, Pelin, Pruesse, Elmar, Glöckner, Frank Oliver, Ludwig, Wolfgang, Schleifer, Karl-Heinz, Whitman, William B, Euzeby, Jean, Amann, Rudolf, & Rossello-Mora, Ramon. 2014. Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nature Reviews Microbiology*, **1**2(9), 635.

Acknowledgements

Diese Arbeit wurde gefördert durch das Bundesministerium für Bildung und Forschung über das Verbundprojekt MÄQNU (BMBF 03MS642A, 03MS642H) und durch die Deutsche Forschungsgemeinschaft (DFG KR 2231/6-1).

Teil V.

Anhang

A. Abstandswahl

Überlegungen

Ziel ist es an dieser Stelle, für den Abstand zwischen den Mittelpunkten von zwei Gruppen X und Y unabhängiger Zufallsvektoren ein geeignetes Maß für einen geeigneten Intervall-Schätzer (beruhend auf paarweisen Abständen) zu entwickeln. Hierbei erscheint zunächst folgende Statistik vielversprechend:

$$r_{bw}(x,y) := \frac{1}{n_x n_y} \sum_{\substack{1\leq i\leq n_x \\ 1\leq j\leq n_y}} d(x_i,y_j) - \frac{1}{2}\left(\frac{1}{n_x^2}\sum_{1\leq i,j\leq n_x} d(x_i,x_j) + \frac{1}{n_y^2}\sum_{1\leq i,j\leq n_y} d(y_i,y_j)\right) .$$

Beispielsweise für den quadratischen euklidischen Abstand $d(a,b) := \sum_{k=1}^{p}(a_k - b_k)^2$:

$$r_{bw}(x,y) := \frac{1}{n_x n_y} \sum_{\substack{1\leq i\leq n_x \\ 1\leq j\leq n_y}} \sum_{k=1}^{p}(x_{ik}-y_{jk})^2 - \frac{1}{2}\left(\frac{1}{n_x^2}\sum_{1\leq i,j\leq n_x}\sum_{k=1}^{p}(x_{ik}-x_{jk})^2 + \frac{1}{n_y^2}\sum_{1\leq i,j\leq n_y}\sum_{k=1}^{p}(y_{ik}-y_{jk})^2\right) .$$

Abstand der Mittelpunkte

Für Abstände, die sich für ein beliebiges reelles Skalarprodukt gemäß

$$d(x,y) = ||x-y||^2 =< x-y, x-y > \ . \tag{A.1}$$

definieren lassen, kann der Abstand der Mittelpunkte von x und y in paarweise Abstände zerlegt werden.

Satz A.1:
Sei V ein reeller Vektorraum, mit Skalarprodukt $< ., . >$ und daraus resultierender Norm $||.||$. Seien $x_1, \ldots, x_{n_x} \in V$ und $y_1, \ldots, x_{n_y} \in V$.

Dann gilt:

$$\frac{1}{n_x n_y} \sum_{\substack{1 \leq i \leq n_x \\ 1 \leq j \leq n_y}} ||x_i - y_j||^2 = ||\bar{x} - \bar{y}||^2 + \frac{1}{n_x^2} \sum_{1 \leq i < j \leq n_x} ||x_i - x_j||^2 + \frac{1}{n_y^2} \sum_{1 \leq i < j \leq n_y} ||y_i - y_j||^2 .$$

Lemma A.2:

$$\frac{1}{n_x} \sum_{i=1}^{n_x} ||x_i - \bar{x}||^2 = \frac{1}{2n_x^2} \sum_{1 \leq i,j \leq n_x} ||x_i - x_j||^2 = \frac{1}{n_x^2} \sum_{1 \leq i < j \leq n_x} ||x_i - x_j||^2$$

Beweis von Lemma A.2.

$$\begin{aligned}
\sum_{i=1}^{n_x} ||x_i - \bar{x}||^2 &= \sum_{i=1}^{n_x} \left(||x_i||^2 - 2 < \bar{x}, x_i > + ||\bar{x}||^2 \right) && = \sum_{i=1}^{n_x} \left(||x_i||^2 \right) - n_x ||\bar{x}||^2 \\
&= \sum_{i=1}^{n_x} \left(||x_i||^2 \right) - \frac{1}{n_x} \sum_{1 \leq i,j \leq n_x} < x_i, x_j > && = \frac{1}{n_x} \sum_{1 \leq i,j \leq n_x} \left(\frac{||x_i||^2 + ||x_j||^2}{2} - < x_i, x_j > \right) \\
&= \frac{1}{2n_x} \sum_{1 \leq i,j \leq n_x} ||x_i - x_j||^2 && = \frac{1}{n_x} \sum_{1 \leq i < j \leq n_x} ||x_i - x_j||^2
\end{aligned}$$

□

Beweis von Satz A.1.

$$\begin{aligned}
\sum_{\substack{1\le i\le n_x\\1\le j\le n_y}} ||x_i - y_j||^2 &= \sum_{\substack{1\le i\le n_x\\1\le j\le n_y}} \big|\big|(x_i-\bar{x}) - (y_j-\bar{y}) + (\bar{x}-\bar{y})\big|\big|^2 \\
&= \sum_{\substack{1\le i\le n_x\\1\le j\le n_y}} \Bigg(||x_i-\bar{x}||^2 + ||y_j-\bar{y}||^2 + ||\bar{x}-\bar{y}||^2 \\
&\qquad\qquad + 2 < x_i-\bar{x}, (\bar{x}-\bar{y}) - (y_j-\bar{y}) > + 2 < y_j-\bar{y}, \bar{x}-\bar{y} > \Bigg) \\
&= \sum_{\substack{1\le i\le n_x\\1\le j\le n_y}} \left(||x_i-\bar{x}||^2 + ||y_j-\bar{y}||^2 + ||\bar{x}-\bar{y}||^2 \right) \\
&= n_x n_y \left(||\bar{x}-\bar{y}||^2 + \frac{1}{n_x}\sum_{i=1}^{n_x} ||x_i-\bar{x}||^2 + \frac{1}{n_y}\sum_{i=1}^{n_y} ||y_i-\bar{y}||^2 \right) \\
&= n_x n_y \left(||\bar{x}-\bar{y}||^2 + \frac{1}{n_x^2}\sum_{1\le i<j\le n_x} ||x_i-x_j||^2 + \frac{1}{n_y^2}\sum_{1\le i<j\le n_y} ||y_i-y_j||^2 \right)
\end{aligned}$$

□

Aus Satz A.1 folgt für die Statistik

$$\begin{aligned}
r_{bw}(x,y) &:= \frac{1}{n_x n_y}\sum_{\substack{1\le i\le n_x\\1\le j\le n_y}} ||x_i-y_j||^2 - \frac{1}{2}\left(\frac{1}{n_x^2}\sum_{1\le i,j\le n_x} ||x_i-x_j||^2 + \frac{1}{n_y^2}\sum_{1\le i,j\le n_y} ||y_i-y_j||^2 \right) \\
&= \frac{1}{n_x n_y}\sum_{\substack{1\le i\le n_x\\1\le j\le n_y}} ||x_i-y_j||^2 - \left(\frac{1}{n_x^2}\sum_{1\le i<j\le n_x} ||x_i-x_j||^2 + \frac{1}{n_y^2}\sum_{1\le i<j\le n_y} ||y_i-y_j||^2 \right)
\end{aligned}$$

direkt:

$$\implies r_{bw}(x,y) = ||\bar{x}-\bar{y}||^2 .$$

Transformation

In einem euklidischen Vektorraum erhält man den euklidischen Abstand daraus durch eine Wurzeltransformation. Eine Verallgemeinerung von Satz A.1 auf andere Abstandsmaße ist im allgemeinen nicht möglich. Wenn die Statistik r_{bw} als Heuristik beibehalten wird, kann

im allgemeinen auch keine Wurzel gezogen werden, da r_{bw} für andere Fälle auch negative Werte annehmen kann. In solch einem Fall lässt sich, anstatt einer Wurzel, die Funktion $r \mapsto sign(r)\sqrt{|r|}$ verwenden, die im positiven Fall gerade der Wurzel entspricht und die Monotonie-Eigenschaften beibehält. Falls stattdessen die Konvexität erhalten bleiben soll, kann $r \mapsto \sqrt{r_+}$ verwendet werden. Hierbei ist $r_+ := \max(0, r)$.

normaler univariater euklidischer Abstand

Wie man an dem Fall $\forall i, j : x_i \geq y_j$ sehen kann, ist eine Statistik der Form $r_{bw} = r_b(X, Y) - (r_b(X, X) - r_b(Y, Y))/2$ nicht unbedingt gut geeignet, um $|\bar{x} - \bar{y}|$ zu erweitern:

$$\frac{1}{n_x n_y} \sum_{\substack{1 \leq i \leq n_x \\ 1 \leq j \leq n_y}} |x_i - y_j| = \frac{1}{n_x n_y} \sum_{\substack{1 \leq i \leq n_x \\ 1 \leq j \leq n_y}} x_i - y_j = \bar{x} - \bar{y} = |\bar{x} - \bar{y}| \,.$$

Der Term $r_b(X, Y)$ aus Gleichung (4.1), Seite 27, könnte im (nicht-quadratischen) euklidischen Fall eine bessere Wahl sein.

Betrachtungen dazu erfolgen hier nur univariat und können als Spezialfall zur Überprüfung der Korrektheit von Algorithmen in Simulationsstudien verwendet werden.

Notation 7. Bezeichne $x_{(i)}$ das i-t-kleinste Element von x und $\text{rank}(x)$ den Vektor der Ränge von x, wobei der Rang 1 an das kleinste Element vergeben wird. Falls mehrere Elemente gleich groß sind, sei deren Anordnung untereinander beliebig und ihr Rang unterschiedlich.

Notation 8. $\mathbf{1}$ bezeichne den Vektor, jeweils geeigneter Dimension, der in jeder Koordinate 1 ist.

Lemma A.3:
Mit Notation 7 gilt:

$$\sum_{1 \leq i < j \leq n_x} |x_i - x_j| = 2x^T \left(\textit{rank}(x) - \frac{n+1}{2} \mathbf{1} \right) . \tag{A.2}$$

Beweis.

$$\begin{aligned}
\sum_{1\le i<j\le n_x} |x_i - x_j| &= \sum_{1\le i<j\le n_x} x_{(j)} - x_{(i)} = \sum_{i=1}^{n_x-1}\sum_{j=i+1}^{n_x} x_{(j)} - x_{(i)} \\
&= \sum_{i=1}^{n_x-1}\left(-(n_x-i)x_{(i)} + \sum_{j=i+1}^{n_x} x_{(j)}\right) \\
&= (n_x-1)x_{(n_x)} + (n_x-3)x_{(n_x-1)} + \ldots + (1-n_x)x_{(1)} \\
&= 2\sum_{i=1}^{n_x} i x_{(i)} - n_x(n_x+1)\bar{x} = 2\left(x^T \text{rank}(x) - \frac{n_x(n_x+1)}{2}\bar{x}\right) \\
&= 2x^T\left(\text{rank}(x) - \frac{n_x+1}{2}\mathbf{1}\right)
\end{aligned}$$

□

$\frac{1}{n_x(n_x-1)}\sum_{1\le i<j\le n_x}|x_i - x_j|$ ist auch als Gini-Mean-Difference bekannt.

Definition A.1. Zu der aus den beiden Stichproben x und y gepoolten Stichprobe $z^T := (x_1, \ldots, x_{n_x}, y_1, \ldots, y_{n_y})$ seien die Funktionen I_x und I_y, die jeweils die Ränge von x bzw. y in dieser gepoolten Stichprobe angeben, folgend definiert:

$$\begin{aligned}
I_x : \{1,\ldots,n_x\} &\to \{1,\ldots,n_x+n_y\} \\
i &\mapsto (I_x)_i := I_x(i), \qquad \text{so dass } x_{(i)} = z_{(I_x(i))} \\
I_y : \{1,\ldots,n_y\} &\to \{1,\ldots,n_x+n_y\} \\
i &\mapsto (I_y)_i := I_y(i), \qquad \text{so dass } y_{(i)} = z_{(I_y(i))}
\end{aligned}$$

und $I_x(\{1,\ldots,n_x\}) \cup I_y(\{1,\ldots,n_y\}) = \{1,\ldots,n_x+n_y\}$.

Auch diese und die nächste Definition sind nicht eindeutig, falls mehrere Werte identisch sind. Sie werden hier aber nur verwendet, wenn die jeweiligen Gleichung eindeutig bleiben, weil die uneindeutigen Terme geeignet gemeinsam auftreten.

$$\Delta\text{rank}(x)_i := I_x(i) - \text{rank}(x)_i$$

Satz A.4:

$$\sum_{\substack{1\le i\le n_x \\ 1\le j\le n_y}} |x_i - y_j| = 2\left(x^T\left(\Delta\textit{rank}(x) - \frac{n_y}{2}\mathbf{1}\right) + y^T\left(\Delta\textit{rank}(y) - \frac{n_x}{2}\mathbf{1}\right)\right) \tag{A.3}$$

Beweis.

$$\begin{aligned}
\sum_{\substack{1\le i\le n_x\\ 1\le j\le n_y}} |x_i - y_j| &= \sum_{1\le i<j\le n_x+n_y} |z_i - z_j| - \sum_{1\le i<j\le n_x} |x_i - x_j| - \sum_{1\le i<j\le n_y} |y_i - y_j| \\
&= 2\Bigg(z^T \text{rank}(z) - \frac{(n_x+n_y+1)(n_x+n_y)}{2} \underbrace{\bar{z}}_{(n_x\bar{x}+n_y\bar{y})/(n_x+n_y)} \\
&\quad - x^T \text{rank}(x) - \frac{(n_x+1)n_x}{2}\bar{x} - y^T \text{rank}(y) - \frac{(n_y+1)n_y}{2}\bar{y} \Bigg) \\
&= 2\Bigg(z^T \text{rank}(z) - \frac{(n_x+n_y+1)}{2}(n_x\bar{x}+n_y\bar{y}) \\
&\quad - x^T \text{rank}(x) - \frac{(n_x+1)n_x}{2}\bar{x} - y^T \text{rank}(y) - \frac{(n_y+1)n_y}{2}\bar{y} \Bigg) \\
&= 2\left(z^T \text{rank}(z) - x^T \text{rank}(x) - y^T \text{rank}y \right) \\
&\quad + ((n_x+1) - (n_x+n_y+1))n_x\bar{x} + (n_y+1-(n_x+n_y+1))n_y\bar{y} \\
&= 2\left(z^T \text{rank}(z) - x^T \text{rank}(x) - y^T \text{rank}y \right) - n_x n_y(\bar{x}+\bar{y}) \\
&= 2\left(\sum_{i=1}^{n_x+n_y} i z_{(i)} - \sum_{i=1}^{n_x} i x_{(i)} - \sum_{i=1}^{n_y} i y_{(i)} - \frac{n_x n_y}{2}(\bar{x}+\bar{y}) \right) \\
&= 2\left(\sum_{i=1}^{n_x} I_x(i) z_{I_x(i)} - \sum_{i=1}^{n_x} i x_{(i)} + \sum_{i=1}^{n_y} I_y(i) z_{I_y(i)} - \sum_{i=1}^{n_y} i y_{(i)} - \frac{n_x n_y}{2}(\bar{x}+\bar{y}) \right) \\
&= 2\left(\sum_{i=1}^{n_x} \underbrace{\Big(I_x(i) - i\Big)}_{\in\{0,\dots,n_y\}} x_{(i)} + \sum_{i=1}^{n_y} \underbrace{\Big(I_y(i) - i\Big)}_{\in\{0,\dots,n_x\}} y_{(i)} - \frac{n_x n_y}{2}(\bar{x}+\bar{y}) \right) \\
&= 2\left(x^T \Delta\text{rank}(x) + y^T \Delta\text{rank}(y) - \frac{n_x n_y}{2}(\bar{x}+\bar{y}) \right) \\
&= 2\left(x^T \left(\Delta\text{rank}(x) - \frac{n_y}{2}\mathbf{1} \right) + y^T \left(\Delta\text{rank}(y) - \frac{n_x}{2}\mathbf{1} \right) \right)
\end{aligned}$$

□

Satz A.5:
Seien $X \sim N(\mu_x, \sigma)$ *und* $Y \sim N(\mu_y, \sigma)$ *unabhängig. Dann gilt:*

$$E|X-Y| = \frac{\sigma}{\sqrt{\pi}} \left(2e^{-\left(\frac{\mu_x-\mu_y}{2\sigma}\right)^2} \right) + (\mu_x - \mu_y) \left(1 - 2\Phi\left(\frac{\mu_y - \mu_x}{\sqrt{2}\sigma} \right) \right). \tag{A.4}$$

Beweis. Seien $Z \sim N(0,1)$ und $c \in \mathbf{R}$. Dann gilt:

$$\begin{aligned}
E|Z+c| =& E\Big((Z+c)\left(\mathbb{1}_{(Z>c)} - \mathbb{1}_{(Z<c)}\right) \Big) \\
=& E\Big((Z+c)\left(1 - 2\mathbb{1}_{(Z<c)}\right) \Big) \\
=& E\Big(Z\left(1 - 2\mathbb{1}_{(Z<c)}\right) + c\left(1 - 2\mathbb{1}_{(Z<c)}\right) \Big) \\
=& \left(-2\int_{-\infty}^{c} \frac{t}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}\, dt \right) + c\left(1 - 2\Phi\left(c\right)\right) \\
=& \left(2\int_{0}^{u(c)} \frac{1}{\sqrt{2\pi}}\, du \right) + c\left(1 - 2\Phi\left(c\right)\right) \\
& \Big(\text{mit:} \quad u = e^{-\frac{t^2}{2}}, \qquad du = -te^{-\frac{t^2}{2}}\, dt\Big) \\
=& \frac{1}{\sqrt{2\pi}} 2e^{\frac{-c^2}{2}} + c\left(1 - 2\Phi\left(c\right)\right) .
\end{aligned}$$

Mit $c = (\mu_x - \mu_y)/(\sqrt{2}\sigma)$ gilt $(X-Y)/(\sqrt{2}\sigma) - c \ \sim N(0,1)$ und somit:

$$\begin{aligned}
E|X-Y| =& E|\sqrt{2}\sigma Z + \sqrt{2}\sigma c| = \sqrt{2}\sigma E|Z+c| \\
=& \frac{\sigma}{\sqrt{\pi}} 2e^{-\frac{((\mu_x-\mu_y)/(\sqrt{2}\sigma))^2}{2}} + (\mu_x - \mu_y)\left(1 - 2\Phi\left((\mu_x-\mu_y)/(\sqrt{2}\sigma)\right)\right) \\
=& \frac{\sigma}{\sqrt{\pi}} 2e^{-\left(\frac{\mu_x-\mu_y}{2\sigma}\right)^2} + (\mu_x - \mu_y)\left(1 - 2\Phi\left(\frac{\mu_x-\mu_y}{\sqrt{2}\sigma}\right)\right)
\end{aligned}$$

□

B. Biologie

Als Grundlage für einige phylogenetische Konzepte wird zunächst der Begriff Baum aus der Graphentheorie benötigt.

Definition B.1 (Baum). Ein "Baum" (bzw. "Tree") besteht aus zwei Mengen für die folgendes gilt: Die Elemente der ersten Mengen werden "Knoten" genannt. Die Elemente der zweiten Menge werden "Kanten" genannt und bestehen jeweils aus einem Paar unterschiedlicher Knoten. Ein n-Tupel von Kanten der Form ($(x_0, x_1), (x_1, x_2), \ldots, (x_{n-1}, x_n)$) wird "Pfad der Länge n von x_0 nach x_n" genannt. Hierbei sind $x_0, \ldots, x_n$ paarweise verschiedene Knoten. Zwischen je zwei verschiedenen Knoten existiert ein eindeutiger Pfad. Dies definiert einen "freien Baum". Wird ein Knoten als "Wurzel" ausgezeichnet, ist der Baum nicht mehr frei. Dies definiert eine Hierarchie mit der Wurzel an der Spitze.
[Knuth (2007a)]

Notation 9. Ist ein Pfad monoton bezüglich der Hierarchie, wird er "Ast" genannt. Jede Kante definiert einen "Vater-" und einen "Kindknoten", wobei das Kind niedriger in der Hierarchie steht. Knoten ohne Kind werden "Blätter" genannt. Alle anderen Knoten sind "innere Knoten". Ein Baum mit Wurzel ist ein "binärer Baum", wenn kein Knoten mehr als zwei Kinder hat. Bei binären Bäumen werden die Kinder eigentlich geordnet (links und rechts) – dies kann und sollte im phylogenetischen Kontext ignoriert werden. Sind die Knoten und Kanten eines Baums Teilmengen eines anderen Baums, wird erster auch als "Unterbaum" bezeichnet.

Bemerkung 33:
Kanten können mit Gewichten versehen werden, die auch Kantenlängen genannt werden. Werden diese ignoriert, spricht man auch von der Topologie des Baums (Tree-Topologie).

B.1. Genetik

Die meisten genetischen Prozesse verlaufen in allen Lebewesen sehr ähnlich.

B.1.1. Grundlegende Begriffe

Da es schwer ist alle Fachbegriffe sinnvoll in einer Reihenfolge einzuführen und zu belegen, werden einige grundlegende Begriffe an dieser Stelle vorweg kurz erklärt:

Protein: ein biologisches Molekül, das aus einer Kette von Aminosäuren besteht. In seinem funktionsfähigen Zustand ist es zu einer räumlichen Struktur gefaltet.

DNA: Speichermolekül der Erbinformation von Lebewesen.

RNA: Molekül, das Erbinformationen schlechter speichert als DNA und weniger geeignet ist Funktionen auszuführen als Proteine. Sie entspricht einer Kette bzw. Sequenz.

Sekundärstruktur: Oberbegriff zu elementaren Faltungstypen von kleineren Abschnitten eines kettenförmigen Moleküls im Raum. Diese werden selbst wieder zueinander räumlich angeordnet.

Enzym: ein biologisches Molekül, das eine chemische Reaktion katalysiert – d.h. erheblich begünstigt.

DNA-Synthese: die Erzeugung von DNA.

DNA-Replikation: DNA-Synthese entsprechend einer DNA-Vorlage.

Transkription: Synthese von RNA entsprechend einer DNA-Vorlage.

Translation: Proteinsynthese entsprechend einer RNA-Vorlage (mRNA).

DNA-Polymerase: ein Enzym, das die DNA-Replikation durchführt.

RNA-Polymerase: ein Enzym, das die Transkription durchführt.

Ribosom: ein zusammenhängender Komplex aus vielen Molekülen, der die Translation durchführt.

Primer: eine kurze DNA-Kette (manchmal auch RNA), die bei einer DNA-Replikation verlängert werden soll.

Chromosom: ein sehr langer DNA-Strang, der Erbinformation des Organismus enthält.

Zelle: grundlegende Funktionseinheit jedes Organismus.

Zellkern: ein Bestandteil von Zellen in einigen Organismen, der von einer porösen Membran umschlossen ist und die Chromosomen enthält.

Eukaryot: Organismus, dessen Zelle oder Zellen einen Zellkern besitzen.

Prokaryot: einzelliger Organismus, der keinen Zellkern besitzt.

RNA-Virus: RNA-Viren, Retro-Viren und Para-Retroviren durchlaufen alle eine Phase, in der ihre Erbinformation als RNA vorliegt. Sie werden in der vorliegenden Arbeit alle als “RNA-Viren” bezeichnet.

Rekombination: DNA-Ketten werden geschnitten und anders wieder verknüpft. Häufig sind zwei unterschiedliche DNA-Stränge beteiligt. In RNA-Viren funktioniert Rekombination anders, führt aber zu dem vergleichbaren Ergebnis für RNA.

Phylogenetik: Lehre, die die genetischen Verwandtschaftsbeziehungen zwischen Lebewesen und die zugrundeliegenden Prozesse untersucht.

Gene-Tree: ein binärer Baum, der die genetische Verwandtschaft verschiedener DNA-Sequenzen desselben Gens wiedergibt.

Phylogenetischer Baum: ein binärer Baum, der die genetischen Verwandtschaft von Organismen zusammenfasst.

Taxonomie: die Einteilung aller Lebewesen in eine hierarchische Klassenstruktur. Im Gegensatz zur Phylogenetik werden genaue Zusammenhänge zwischen Organismen zugunsten einer einfachen, standardisierten Einteilung vernachlässigt. Die Bäume zur Darstellung der Zusammenhänge sind nicht binär und alle Kanten gleich lang. Hierdurch entstehen Ebenen, die senkrecht zu den Ästen verlaufen.

Gattung: eine der untersten Ebenen in der Taxonomie.

Art: (bzw. Spezie) die taxonomische Ebene unterhalb der Gattung. Gattung und Art werden zusammen als Namen für Organismen verwendet. Konventionell werden sie dann in Schrägdruck gesetzt, wobei abgesehen vom Anfangsbuchstaben der Gattung alle Buchstaben klein geschrieben werden. Häufig wird dabei die Gattung abgekürzt, indem nur der erste Buchstabe geschrieben und ein Punkt hinter ihm gesetzt wird (z.B.: *E. coli*). Einzelne Gattungen und Arten in den Grundlagen-Abschnitten werden nur als Beispiele verwendet, weil Zahlenwerte oder Eigenschaften für sie bekannt sind.

OTU: (Operational Taxonomic Unit) eine willkürlich gesetzte taxonomische Einheit, die keiner klassischen taxonomischen Einteilung folgen muss. ((Gelegentlich wird der Begriff auch so interpretiert, dass er sich speziell auf die Klassifikation mit dem häufig benutzten Cut-Off-Wert 97% Sequenzübereinstimmung für die hierarchische Clusterung bezieht.))

Alignment: das Schema, das entsteht, wenn zwei oder mehr Sequenzen untereinander geschrieben werden, so dass die einzelnen Positionen möglichst übereinstimmen. Hierfür können Lücken in einzelnen Sequenzen eingefügt werden.

Codon: drei aufeinander-folgende RNA-Basen in einem Leseraster der Translation. ((Die entsprechenden DNA-Basen werden gelegentlich ebenfalls Codon genannt.))

lateraler Gentransfer: Transfer von genetischem Material zwischen Organismen der nicht durch Vererbung geschieht. Anschließend können diese Gene auch auf nachkommende Generationen vererbt werden.

DNA-Sequenzierung: das Bestimmen der Sequenz einer DNA.

DNA-Amplifikation: das Vervielfältigen von DNA-Fragmenten.

PCR: Polymerase Chain Reaction - das aktuell labortechnische Standardverfahren zur Amplifikation von DNA.

Marker-Gen: ((DNA-Sequenzabschnitt, der für Zuordnungen, z.B. des Lebewesens, benutzt werden kann.))

Mutation: eine Änderung an einer genetischen Sequenz. ((Wenn Mutationen einen Vorteil erzeugen sind sie "positiv". Erzeugen sie einen Nachteil, sind sie "negativ". Andernfalls sind sie "neutral".))

akzeptierte Mutation: eine Mutation, die nicht im Laufe der Evolution verloren ging.

Mutationsrate: die Häufigkeit genetischer Änderungen pro ursprünglicher Base pro Generation. Für Viren wird ein Generationszyklus in der Regel mit dem Zellzyklus der infizierten Zelle gleichgesetzt.

[Campbell *et al.* (1999), Lewin (2000), Schulz & Schrimer (1979), Caporaso *et al.* (2010), Cormen *et al.* (2001), Rokas *et al.* (2018), Sanjuan & Domingo-Calap (2016)]

B.1.2. DNA und RNA

B.1.2.1. Nukleinbasen

Sowohl DNA- als auch RNA-Stränge bestehen hauptsächlich aus einer gerichteten Kette von Nukleotiden. Nukleotide bestehen aus drei Komponenten: einem Zuckermolekül, aus einer Base und einem Phosphatrest. Von allen drei Komponenten existieren mehrere Varianten. Der Zucker bestimmt, ob es sich um einen Baustein eines DNA- oder RNA-Strangs handelt. Die Base bestimmt die Erbinformation. Der Phosphatrest kann aus einer Kette von einem, zwei oder drei Phosphatgruppen bestehen. Sowohl in DNA- als auch in RNA-Strängen, kommt nur eine Phosphatgruppe vor. Wenn nur der Phosphat-freie Anteil eines Nukleotids gemeint ist, spricht man von einem Nukleosid. Es gibt vier Arten von Nukleosidmonophosphaten, die in DNA vorkommen:

dAMP Desoxyadenosinmonophosphat, mit Base Adenin (A) bzw. Nukleosid Adenosin.

dGMP Desoxyguanosinmonophosphat, mit Base Guanin (G) bzw. Nukleosid Guanosin.

dCMP Desoxycytidinmonophosphat, mit Base Cytosin (C) bzw. Nukleosid Cytidin.

dTMP Desoxythymidinmonophosphat, mit Base Thymin (T) bzw. Nukleosid Thymidin.

Das "P" am Ende steht für "Phosphat" und der Buchstabe davor, für die Anzahl der Phosphatgruppen. "M" bedeutet 1 (Mono), "D" 2 (Di) und "T" 3 (Tri). Das "d" am Anfang steht für die Art des Zuckermoleküls (hier: Deoxyribose) und kennzeichnet hierdurch den Unterschied zu RNA, die Ribose als Zuckermolekül enthält ((wird aber oft weggelassen)). Wenn die Base für einen Zusammenhang keine Bedeutung hat, wird in der Abkürzung häufig ein "N" anstelle des Buchstabens für die Base verwendet (z.B. NTP).

Weil dem Zuckermolekül der DNA eine reaktive OH-Gruppe fehlt ("Des-Oxy"), eignet

es sich besser zur Datenspeicherung RNA. RNA kann dafür auch Funktion ausführen und nimmt damit auf gewisse Weise eine Zwischenstellung zwischen DNA und Proteinen ein. In einigen Viren ist RNA der Träger der Erbinformation. In RNA wird anstatt Thymidin (T) die Base Uracil (U) verwendet (Nukleosid Uridin).
[Wehner & Gehring (1993), Campbell *et al.* (1999), Kaufmann & A. (1996)]

Bemerkung 34:
Abwandlungen der vier Nukleosidtriphosphate werden bei den meisten Sequenzierverfahren eingesetzt. Die Phosphatgruppen werden bei der Pyrosequenzierung ausgenutzt.

Methylierung: An Nukleinbasen kann eine Methylgruppe angebaut werden. Dies betrifft hauptsächlich Cytosin. Der Prozess wird Methylierung genannt und ist meist mit der Deaktivierung des betroffenen Gens assoziert. Methylgruppen spielen auch eine Rolle bei der Korrektur von Sequenzfehlern. Unmittelbar nach einer DNA-Replikation fehlen dem neuen Strang Methylreste an geeigneten Basen. Ist dies an bestimmten kurzen Sequenzmustern wie "GTAC" der Fall, dann könner Reparaturenzyme den kopierten Strang von dem Original unterscheiden und reparieren.
[Campbell *et al.* (1999), Sanjuan & Domingo-Calap (2016)]

Basenpaare: Die vier Basen lassen sich in zwei Gruppen einteilen: Die beiden Purine Adenin und Guanin haben eine leicht komplexere Struktur und sind daher etwas größer als die beiden Pyrimidine Cytosin und Thymin. Guanin und Cytosin können zwischeneinander drei Wasserstoffbrücken bilden, zwischen Adenin und Thymin können zwei gebildet werden. Wasserstoffbrücken sind relativ schwache Bindungen zwischen Molekülen. Die Wasserstoffbrücken stabilisieren DNA-Stränge, die sich aneinander lagern. Ein Basenpaar besteht aus einem Purin eines DNA-Strangs und dem Wasserstoffbrücken-gebundenen Pyrimidin des komplementären DNA-Strangs. ((Basenpaare werden auch als Einheit, "bp", für die Anzahl der Stellen einer Sequenz verwendet. 1 bp entspricht dabei einer Stelle. Diese Einheit werde ich auch verwenden, wenn keine Paare vorliegen.))
[Wehner & Gehring (1993), Campbell *et al.* (1999), Kaufmann & A. (1996)]

Höhere Guanin-Anteile erleichtern die oxidative Schädigung eines Sequenzabschnitts [Graw (2015)].

Bemerkung 35:
Bereits ohne Evolutionsdruck zeigt sich schon an dieser Stelle, dass nicht alle Mutationen gleichwahrscheinlich sind und beeinflußt werden von:

Base: Guanin erleichter einige Mutationen seiner Umgebung.

Sequenz: einige werden besser repariert, weil sie ein geeignetes Muster aufweisen.

Speichermedium: RNA mutiert leichter.

B.1.2.2. Doppelhelix

Zwei DNA-Einzelstränge können sich zu einem stabilen Doppelstrang in Form einer Doppelhelix zusammenlagern, wenn die Basenpaare des einen Strangs komplementär zu denen des anderen Strangs sind (d.h. zueinander passen – A zu T und G zu C). Die Basen liegen dabei innerhalb der Helix. Diese Struktur ähnelt einer Wendeltreppe, wobei jede Stufe aus zwei Basen, je einer aus jedem der beiden Stränge, gebildet wird. Die G–C Paarung ist stabiler als die A–T, was unter anderem für die Temperaturbeständigkeit von DNA relevant ist. Bei Säugetieren sind ca. 40% der Paarungen G–C. Eine Mutation von einem Pyrimidin in ein Purin oder umgekehrt, stellt, aufgrund der Größenverhältnisse, einen Störfaktor für die Doppelhelix da. Daher kann ein Purin leichter zu einem anderen Purin mutieren als in ein Pyrimidin und ein Pyrimidin leichter zu einem anderen Pyrimidin als zu einem Purin.

Jeder Einzelstrang hat eine Richtung in die er abgelesen werden kann. In einem Doppelstrang sind die Richtungen der Einzelstränge einander immer entgegengesetzt. In Zellen liegen Enzyme vor, die, zu einem DNA-Einzelstrang einen zweiten komplementären Strang synthetisieren. Die Sequenz der Basenfolge eines DNA-Einzelstrangs enthält den Großteil der Information der DNA. Die Sequenz eines Strangs der Helix ist durch den anderen vorgegeben.

Zwei bereits existierende DNA-Fragmente können sich aneinanderlagern und eine Doppelhelix bilden. Die Anlagerung von zwei DNA-Strängen oder eines RNA- an einen DNA-Strang wird Hybridisierung genannt.
[Lewin (2000), Campbell *et al.* (1999), Kaufmann & A. (1996), Graw (2015)]

Bemerkung 36:
Durch Hybridisierungen können Sequenzen "herausgefischt" werden. Sie bilden auch die Startpunkte für Sequenzierungen. Die Konstrukt eines komplementären Strangs, bildet die Grundlage der meisten Sequenzierungsverfahren.

B.1.2.3. Lokal abweichende DNA-Struktur

Besondere Basenfolgen beeinflussen Form und Stabilität der DNA-Doppelhelix. Dies ist teilweise auch mit Funktionen wie Regulation von Replikation und Rekombination verbunden. Die Anfangsstellen der DNA-Replikation sind beispielsweise aufgrund ihres hohen Anteils an "A" und "T" instabil. Ebenso erlauben einige Sequenzen die Bindung von Molekülen, die die Struktur der DNA kurzzeitig verändern. Bei der Transkription ist dies oft der Fall. DNA-Sequenzen können auch dazu führen, dass sich in einem Doppelstrang ein Einzelstrang lokal abtrennt und eine Schleife bildet, die um sich selbst gedreht ist, bevor er mit dem zweiten Strang wieder zusammenläuft.
[Graw (2015)]

Bemerkung 37:
Die Struktur beeinflusst ebenfalls, ob schädigende Substanzen an einer Stelle in Kontakt mit den innen-liegenden Basen kommen können. Die Schleifen können zu Veränderungen führen, die die ganze daran beteiligte Sequenz betrifft oder auch Störungen beim Ablauf der Kopierung der DNA verursachen, die sogar zum Wechsel der Vorlage führen können.

Tabelle B.1.: Genomgrößen aus Graw (2015)

Lebewesen	Basenpaare*
Viren[1]	10^3 - 10^5
Prokaryoten	10^6 - 10^7
Protisten	10^6 - 10^{11}
Pilze	10^7 - 10^9
Pflanze	10^8 - 10^{11}
Säugetiere	10^9 - 10^{10}

Von gleichartigen Chromosomen wurde jeweils nur eins gezählt. *: Die Größen sind nur grob angegeben.
1: Viren zählen nicht als Lebewesen und sind hier nur zusätzlich angegeben.

B.1.2.4. Vorkommen von DNA

Der Hauptanteil der DNA eines Organismus liegt auf einem oder mehreren großen DNA-Doppelsträngen, die Chromosomen genannt werden. Die Größe von Genomen unterscheidet sich zwischen Lebewesen (Tab. B.1). Einige Organismen können zusätzlich kleinere geschlossen DNA-Ringe enthalten, die Plasmide genannt werden. Diese können zwischen Organismen übertragen werden. Viren bestehen entweder aus DNA oder RNA und sind ebenfalls übertragbar. Eukaryoten enthalten einige membranumschlossene Strukturen in ihren Zellen von denen einige eigene Chromosomen besitzen.
[Wehner & Gehring (1993)]

Bemerkung 38:
Zellen können mehrere gleichartige Chromosom aufweisen. Prokaryoten haben typischerweise nur ein Chromosom. Die meisten Pilzzellen haben keine zwei gleichartigen Chromosome [Dutheil (2020)].

B.1.2.5. Replikation

Die Synthese von DNA gemäß einer Vorlage wird Replikation genannt. Sie findet üblicherweise vor einer Zellteilung statt, damit die beiden entstehenden Zellen jeweils das vollständige Genom besitzen können. Sie beginnt normalerweise an einem sogenannten Ursprungspunkt auf dem Chromosom. An einem Ursprungspunkt der Replikation werden die Stränge getrennt und ein kurzer RNA-Primer von 20 bis 500 Basenpaaren gebildet. Eine DNA-Polymerase verlängert diesen Primer durch eine DNA-Sequenz, die zu dem betreffenden Einzelstrang komplementär ist. Diese Polymerase kann den neuen DNA-Strang nur in eine Richtung bilden. Dennoch läuft die Synthese global gesehen an beiden alten Strängen in beide Richtungen ab. (Ausnahmen sind Viren und Plasmide – bei diesen werden Stränge vollständig separat repliziert [Graw (2015)].) Dies geschieht indem eine der Richtungen jeweils Stückweise (ca. 1000 Basenpaare) synthetisiert wird. Die Stücke werden anschließend verbunden. Die DNA wird sofort korrekturgelesen und Fehler werden beseitigt. Insbesondere wird der RNA-Primer durch entsprechende Desoxiribonukleotide ersetzt.

Beim Einbau eines Nukleotids in einen DNA Strang, muss dieses als Desoxyribonukleosidtriphosphat (dNTP) vorliegen. Von diesem wird ein Pyrophosphat ((d.h. eine Kette aus zwei Phosphaten)) abgespalten.
[Hirsch-Kauffmann & Schweiger (1992)]

Replikationsfehler: Ohne Korrektur würde die Mutationsrate durch Replikationsfehler zwischen 10^{-6} und 10^{-5} liegen. Durch Reparaturmechanismen sinkt sie auf 10^{-11} bis 10^{-9}. Die Reparaturmechanismen hängen auch vom Organismus ab. Pathogene Bakterien haben teilweise weniger Gene für DNA-Reparatur und Rekombination als andere Bakterien.
[Graw (2015)]

B.1.2.6. Rekombination

Bei der Rekombination werden DNA-Ketten geschnitten und anders wieder verknüpft. Häufig sind zwei unterschiedliche DNA-Stränge beteiligt. Rekombination und DNA-Reparatur basieren meist auf denselben Mechanismen. Einige DNA-Sequenzen begünstigen Rekombinationsereignisse.
[Graw (2015)]

Homologe Rekombination ist ein Mechanismus, der DNA-Schäden mit hoher Genauigkeit reparieren kann. Er benötigt allerdings lange Bereiche übereinstimmender DNA-Sequenz (ca. 50 bp in *E. coli* und 300 bp in Säugetierzellen) mit der Vorlage. Wenn die Vorlage an einer anderen Position liegt als die zu reparierende Stelle, spricht man von nicht-allelischer homologer Rekombination (NAHR). In diesem Fall sind auch bei homologen Rekombinationen strukturelle Änderungen des Chromosoms möglich.
[Hastings *et al.* (2009)]

Nicht-homologe Rekombination: Prozesse, die keine oder nur geringe Übereinstimmung benötigen, werden als nicht-homologe Rekombination bezeichnet und erzielen meist weniger genaue Reparaturen. Zellulärer Stress führt zu einem höheren Anteil nicht-homologer Reparaturen.
[Hastings *et al.* (2009)]

Rekombination außerhalb von Reparaturprozessen: Für die Integration von Virus-DNA in bakterielle Chromosome reichen wenige identische Basenpaare in beiden Sequenzen aus. Bei transponierenden Elementen (Kapitel B.1.2.7, Seite 115) brauchen die Sequenzen nicht identisch zu sein. Hier liegen passende Enzyme für das Ausschneiden der Sequenz des Transposons vor. Diese kann relativ unspezifisch wieder eingebaut werden. Bei Eukaryoten finden Rekombinationen insbesondere bei geschlechtlichen Teilungen statt, bei der Sequenzen zwischen väterlichen und mütterlichen Chromosomen ausgetauscht werden.
[Graw (2015)]

Bedeutung von Rekombination: Rekombinationen können Untereinheiten von Proteinen neuzusammensetzen und auf diese Weise quasi nach Baukastensystem neue Funktionen schaffen. Je nach Mechanismus kann Rekombination genetische Diversität senken oder erhöhen.
[Hastings *et al.* (2009)]

Außerdem können positive von negativen Mutationen getrennt und mehrere positive Mutationen zusammengefügt werden. Rekombinationen können bis zu einem gewissen Punkt die Anzahl tolerierbarer Mutationen erhöhen. Darüber hinaus wirken dann auch Rekombinationen negativ.
[Sanjuan & Domingo-Calap (2016)]

Bemerkung 39:
Rekombinationen führen zu komplizierten Verwandtschaftsbeziehungen und Zusammenhängen – vor allem in längeren DNA-Sequenzen.

B.1.2.7. Transponierende Elemente

Transponierende (genetische) Elemente sind DNA-Sequenzen, die ihren Platz im Genom und zwischen den Genomen von Lebewesen leicht wechseln können. Die Mechanismen, durch die dies geschieht, unterscheiden sich zwischen den Elementen. Bei einigen Mechanismen bleibt das Gen an seinem alten Ort erhalten, so dass sich seine Anzahl erhöht. Ca. 43% des menschlichen Genoms besteht aus mobiler DNA [Storch *et al.* (2013)]. Sie sind eine Hauptursache für Mutationen. In Prokaryoten wird der Anteil durch diese Elemente entstehender Mutationen auf 20 bis 40% geschätzt, obwohl ein Ortswechsel der Sequenz nur ungefähr einmal in Millionen Generationen auftritt und die Anzahl der Kopien jedes Elements nur gering ist. Die Mutationshäufigkeiten in der Nähe beweglicher Elemente steigt gelegentlich um zwei bis drei Größenordnungen an.
[Wehner & Gehring (1993), Lewin (2000), Graw (2015)]

Es existieren spezielle Genomabwehrmechanismen, die in wiederholte Sequenzen gezielt Punktmutationen einführen und auf diese Weise Transponierende Elemente in ihrer Ausbreitung behindern. Rekombination, insbesondere durch sexuelle Vermehrung, behindern ebenfalls die Ausbreitung.
[Dutheil (2020)]

Spuren von mobilen Elementen finden sich in praktisch allen Eukaryoten. Diese können für Taxen charakteristisch sein und sich in einem bestimmten Zeitraum ausgebreitet haben (Evo277).
[Storch *et al.* (2013)]

Insertionssequenzen stellen die einfachste Form dar. Sie sind zwischen 768 und 2132 Basenpaare lang und existieren in bakteriellen Chromosomen, Bakteriophagen ((d.h. Viren, die

auf Bakterien spezialisiert sind)) und Plasmiden. Sie enthalten sogenannte Erkennungssequenzen an ihren Enden. Dazwischen kodieren sie für Enzyme, die DNA an diesen Erkennungssequenzen zerschneidet. Nach Schnitten an diesen Stellen, kann die Insertionssequenz an anderer Stelle wieder in das Genom eingebaut werden. Am Einbauort reicht eine teilweise Sequenzhomologie ((d.h. Ähnlichkeit der Sequenzen an einigen Stellen)).
[Wehner & Gehring (1993), Lewin (2000), Graw (2015)]

Transposonen sind größere transponierende Elemente (mehrere tausend Basenpaare), die zusätzlich weitere Gene enthalten. Sie können durch Umwelteinflüsse (z.B. Stress) dazu angeregt werden, ihre Position zu wechseln. Bei einem Ortswechsel können benachbarte Sequenzbereiche mittransferiert werden. Ca. 3% des menschlichen Genoms besteht aus Transposons, von denen allerdings keins mehr aktiv ist [Storch *et al.* (2013)]. Eukaryotische Zellen können mehrere hunderttausend (nicht-exakte) Kopien eines Transposons enthalten. Die meisten dieser Kopien sind nicht mehr fähig ihre Position selbstständig zu wechseln. Von medizinischer Bedeutung sind insbesondere Transposonen die Resistenzgenen gegen Antibiotika enthalten. Mehrerer dieser Transposonen liegen oft auf einem Plasmid, dann Resistenzfaktorplasmid (R-Plasmid) genannt, und können zwischen Bakterien übertragen werden. Pathogene Bakterien erhalten durch R-Plasmide einen Selektionsvorteil.
[Lewin (2000), Graw (2015)]

Bemerkung 40:
Medizinisch relevante Eigenschaften werden also häufig zwischen Bakterien übertragen, die in einem phylogenetischen Baum weit auseinanderliegen können.

SINEs: 13% unseres Genoms besteht aus Short Interspersed Elements (SINEs) - diese sind wenige 100bp lang und enthalten Promotorelemente aber keine proteinkodierenden Bereiche. Nach einer Translation wird ihre RNA an andere Stelle wieder in DNA übersetzt und eingebaut, sofern ein Enzym in der Zelle vorhanden ist, das dazu in der Lage ist.
[Storch *et al.* (2013)]

LINEs: 17% underes Genoms besteht aus Long Interspersed Elements (LINEs). Diese sind mehrere kbp lang. Die meisten sind unvollständig. Vollständige LINEs sind ca. 6kbp lang und enthalten zusätzlich zum Promotor Code für eine reverse Transkriptase und Endonuklease, die RNA in DNA übersetzt und letztere scheiden können. Die Vermehrung erfolgt wie bei SINEs. Der Einbau ist nicht gut, weshalb oft nur der Anfang eingebaut wird. SINEs profitieren von LINEs. SINEs und LINEs sind noch heute in Menschen aktiv und vermehren sich.
[Storch *et al.* (2013)]

Plasmide und Viren können selbst transponierende Elemente sein, sich also in das Wirtsgenom integrieren.

Viren enthalten Information für Kapseln, die ihre Erbinformation bei Übergang von einer Zelle in eine möglicherweise weit entfernte Zelle schützen und die Infektion erleichtern. Hüllen sind die Wirts-Membranen, mit denen mache Viren ((z.B. Retroviren und Coronaviren)) ihre Kapseln umgeben. Bei Retro-Viren besteht die Erbinformation aus RNA, die durch Enzyme des Virus in DNA übersetzt wird. Nachdem diese DNA in ein Chromosom des Wirtes eingebaut ist, heißt sie Pro-Virus. Auf diese Weise wird das Virus auch durch Zellteilungen vererbt. Ein vererbter Virus kann manchmal fremde Arten besser infizieren als die Art seines Wirts. Die DNA des Wirts, die direkt hinter dem Pro-Virus liegt, kann in den Retro-Virus integriert werden. Retro-Viren infizieren Eukaryoten.
[Wehner & Gehring (1993), Lewin (2000), Graw (2015)]

Proviren sind ca. 9kbp lang. 8% des menschlichen Genoms sind ehemalige Proviren von Retroviren. Einige davon sind noch relativ intakt.
[Storch *et al.* (2013)]

B.1.3. Proteinsynthese

Proteinsynthese (die Herstellung von Proteinen) verläuft im wesentlichen in zwei Schritten.

Transkription:
: Eine RNA-Polymerase lagert sich an einen besonderen DNA-Abschnitt an, der Promotor genannt wird. Die Basensequenz auf dem Strang wird von ihr ausgelesen und in eine sogenannte mRNA (Messenger-Ribonukleinsäure) mit entsprechender Sequenz synthetisiert – bei Eukaryoten wird dieses RNA-Molekül zunächst prä-mRNA genannt, bis es aus dem Zellkern transportiert und beim sogenannten Splicing nochmals verändert wird. Die beiden Sequenzen (DNA und mRNA bzw. prä-mRNA in Eukaryoten) haben die gleiche Länge und sind eineindeutig einander zugeordnet. Die Synthese der mRNA endet, nachdem die RNA-Polymerase auf einen weiteren besonderen DNA-Abschnitt (Terminator) trifft. Den Abschnitt zwischen Promotor und Terminator nennt man Gen. ((Es gibt abweichende Definitionen von "Gen".))

Translation:
: Die Proteinsynthese beginnt an einer speziellen Abfolge von drei Basen der mRNA, die Startcodon genannt wird. Es bildet sich ein sogenannter Initiationskomplex aus mRNA, sogenannter Initiator-tRNA und der kleinen Untereinheit eines Ribosoms. tRNAs sind spezielle RNA-Moleküle, an denen eine zugehörige Aminosäure gebunden ist. Die Initiator-tRNA bindet an das Startcodon. Anschließend wird durch Bindung der großen Untereinheit das Ribosom vervollständigt. Das Ribosom benutzt ein Leseraster von jeweils drei Basen der mRNA. Diese drei Basen werden Codon genannt. Nachdem diese Basen gelesen wurden, wird die Position um drei Basen verschoben. Jedes Codon wird somit in eine Aminosäure übersetzt. Dies geschieht indem sich eine passende tRNA (transfer Ribonukleinsäure) an das Ribosom anlagert und die an ihr gebundene Aminosäure

an eine entstehende Kette anbaut. Das Ribosom kann zwei tRNAs gleichzeitig binden. Die erste gehört zu der bereits gebildeten Kette und wird von dieser abgestoßen, sobald es die Bindungsstelle des Ribosoms verlässt. Die zweite gehört zu der nächsten Aminosäure, die eingebaut werden soll. Das Leseraster wird dadurch verschoben, dass die zweite tRNA in die Position der ersten vorrückt. Die Translation endet wenn das Ribosom auf eines von drei möglichen Stopcodons (UAA, UAG oder UGA) trifft.

[Knippers (1997), Wehner & Gehring (1993), Graw (2015)]

Die entstandene Aminosäurekette wird gegebenenfalls noch an weitere Aminosäureketten angelagert, verändert, in eine dreidimensionale Struktur überführt und ist dann ein fertiges Protein. Bei der Translation kodieren drei Stellen aus jeweils vier möglichen Basen ($4^3 = 64$) für 20 Aminosäuren. Für die meisten Aminosäuren existieren mehrere Codons. Der Code ist für alle Organismen gleich und wird "genetischer Code" genannt.
[Knippers (1997), Wehner & Gehring (1993)]

Teilweise wird die Reduktion von 64 Basenkombinationen auf 20 Aminosäuren dadurch umgesetzt, dass einige tRNAs an der dritten Position des Codons nicht vollkommen spezifisch sind, sondern dort verschiedene Basen akzeptieren. Dies ist als “Wobble-Hypothese” bekannt. Auch die erste Position im genetischen Code kann (wenn auch wesentlich seltener) verschiedene Basen aufweisen, die zur derselben Aminosäure führen. Die zweite Position ist hingegen für jede Aminosäure eindeutig.
[Wehner & Gehring (1993)]

Notation 10. Wenn der Fokus der Proteinsynthese nicht auf dem Protein oder dem Prozess liegt, sondern auf den Genen, wird auch von (Gen-)Expression gesprochen. Man sagt ein Gen “werde exprimiert” oder “sei aktiv”.

Die Konzentration benötigter tRNAs in der Zelle beeinflusst die Geschwindigkeit maßgeblich, mit der eine Expression abläuft. Das hat auch Auswirkungen auf Mutationen. Chromosome sind so aufgebaut, dass Condons für verfügbarere tRNAs häufiger vorkommen. Diese relativen Häufigkeiten unterscheiden sich zwischen Organismen.
[Graw (2015)]

Bemerkung 41:
Ribosomen bilden die Grundlage vieler Mikrobiomanalysen und die Eigenarten der Translation tragen Bedeutung für den Zusammenhang zwischen Evolution und Sequenz.

Bemerkung 42:
Der Begriff “Gen” ist im Grunde analog zu dem Begriff “Computer-Datei” zu verstehen. Das Gen ist gewissermaßen der virtuelle Ort, an dem bestimmte Information zu finden ist, oder ihr Zweck. Im Computer liegt die Datei in einer 0-1-kodierten Zeichenkette, die physikalisch in weiter angeordneten Strukturen (z.B. Ringe einer Festplatte) steckt. Im Lebewesen ist das Gen in einer A–C–G–T-kodierten Zeichenkette, die ebenfalls physikalisch komplizierter angeordnet ist. So wie Dateien nicht in der selben Reihenfolge auf einer Festplatte gespeichert zu sein brauchen, können auch Gene verschieden angeordnet sein. Bei verwandten Organismen sind sie jedoch

weitgehend identisch angeordnet. Aussagen darüber, dass Menschen und Affen x% gemeinsame Gene haben, entsprechen ungefähr der, dass der Computer in meinem Büro und der zu Hause x% gemeinsame Konfigurationsdateien haben. Dass sich deren Inhalte (und somit die Einstellungen) unterscheiden, ist dabei nicht berücksichtigt. Wenn sich die Sequenzen eines Gens in zwei Lebewesen unterscheiden, spricht man von zwei "Allelen" desselben Gens und sagt, dass sich die Lebewesen in ihrem "Genotyp" unterscheiden. Der Genotyp beschränkt sich nicht auf ein Gen, sondern erstreckt sich über die Gesamtheit der betrachteten genetischen Information. Auf eine saubere Benutzung des Worts "Gen" wird häufig genauso wenig Wert gelegt wie bei dem Wort "Datei".

B.1.3.1. Heterochromatin

Eukaryotische Chromosomabschnitte werden aufgeteilt in Heterochromatin und Euchromatin. Heterochromatin ist dichter gepackte DNA, die schlechter zugänglich ist und meist nicht exprimiert wird. Das kann sich auch auf Gene auswirken, wenn sie ihre Position im Chromosom ändern. Es sind aber auch einige wenige Gene bekannt, die nur im Hetorochromatin exprimiert werden. Heterochromatin enthält einen viel höheren Anteil an repetitiven Sequenzen. Die Enden von eukaryotischen Chromosomen gehören zum Heterochromatin und werden "Telomere" genannt.
[Graw (2015)]
Im Gegensatz zu vererbter DNA werden Telomere durch Telomerasen (spezielle Enzyme) an Chromosomenden angeheftet.
[Storch *et al.* (2013)]

Bemerkung 43:
Heterochromatin ist weniger anfällig für mutationserregende Substanzen und unterliegt seltener Selektionsdruck.

B.1.4. Mutationen

Mutationen sind Veränderungen der Erbinformation. Traditionell wird unterschieden in:

Genom-Mutation: eine drastische Veränderung wie die Erhöhung der Anzahl eines ganzen Chromosoms. Bei lebensfähigen Menschen kann dies beispielsweise bei Chromosom 21 (Trisomie 21) und den Sexchromsomen auftreten.

Translokation: Verlagerung der Position einer Sequenz im Genom (siehe transponierende Elemente, Seite 115).

Deletion: ein oder mehrere Basenpaare gehen verloren.

Veränderung der Kopienzahl: die Häufigkeit, mit der eine Sequenz im Genom vorkommt, wird verändert. Dies führt zu Kopienzahlvariationen (CNV) zwischen Genomen.

Insertion: ein oder mehreren Basenpaare werde neu eingebaut.

Inversion: die Sequenz eines DNA-Abschnitts wird umgedreht. (Das Ende des Abschnitts wird zum Anfang und der Anfang zum Ende.)

Punktmutation: bzw. Nukleotid-Austausch. Eine Base wird gegen eine andere ausgetauscht. Wenn dies in einem Organismus geschehen ist, besteht für die Population dieser Organismen an dieser Stelle ein sogenannter "Single Nucleotid Polymorphism" (SNP). ((Manchmal werden auch Mutationen, die eine einzelne Base betreffen, Punktmutationen genannt. Die Deletion einer Base wäre dann auch eine Punktmutation.))

[Knippers (1997), Hein *et al.* (2005), Graw (2015)]

Bemerkung 44:
In einem Vergleich zwischen einem humanen Genom und dem Humanen-Referenz-Genom lagen nur an 0.1% der vergleichbaren Stellen SNPs vor. Hingegen lagen 1.1% weitere Abweichungen zwar an Mutationen, aber nicht an Punktmutationen.
[Dutheil (2020)]

Bemerkung 45:
Einige Mutationen geschehen seltener, weil sie z.B. die Struktur der DNA-Doppelhelix stören oder leichter reparierbar sind. Andere Mutationen sind seltener in heutigen Lebewesen zu beobachten, weil die betroffenen Vorfahren ausgestorben sind. Für die Zwecke der vorliegenden Arbeit wird nicht strickt zwischen Mutationen und sogenannten akzeptierten Mutationen (also nur die, die sich durchgesetzt haben) unterschieden. In dem Sinne hat Selektion einen Einfluss auf die Mutationswahrscheinlichkeit, auch wenn der Mutations-erzeugende Prozess nicht beeinflusst wurde.

B.1.4.1. Punktmutationen

Punktmutationen, im weiteren Sinne, entstehen durch Desaminierung, Depurinierung, Dimerisierung oder Oxydierung.
[Storch *et al.* (2013)]

Transition: Als "Transition" wird eine Mutation bezeichnet, wenn ein Purin zu einem anderen Purin, bzw. Pyrimidin zu einem anderen Pyrimidin wird. Über 90% der SNPs in homologen Genen zwischen nah-verwandten Arten sind Transitionen. Zwischen entfernt-verwandten sinkt der Anteil bis auf 35%, wegen multipler Mutationen. Alle anderen Punktmutationen sind "Transversionen" – ausgenommen Deletionen und Insertionen.
[Storch *et al.* (2013)]

Desaminierung: In Eukaryoten finden 100 Desaminierungen pro Tag und Zelle statt. Bei einer Desaminierung wird entweder ein (eventuell methyliertes) Cytosin zu Uracil und im weiteren Verlauf der Vererbung zu Thymin (Transition) oder ein Adenin zu Hypoxanthin ((und dann zu Guanin – eine Transversion)).
[Storch *et al.* (2013)]

Oxidation: Guanin wird leicht oxidiert und dabei zu 8-Oxoguanin und im weiteren Verlauf der Vererbung zu Thymin (Transversion). Oxidation ist maßgeblich für Alterung verantwortlich.
[Storch *et al.* (2013)]

Depurinierungen sind die Entfernungen von Purinen (A, G) und führen zu Deletionen oder willkürlichem Baseneinbau. Mit 5000 bis 10000 Depurinierungen im Mensch pro Tag und Zelle gehören sie zu den häufigsten Mutationen. Sie können Einzell- und Doppelstrangbrüche verursachen.
[Storch *et al.* (2013)]

Dimerisierung: UV-Licht bewirkt, dass aus CC oder TT ein Dimer wird. Im weiteren Verlauf der Vererbung entsteht eine Deletion. Hauptsächlich ist TT betroffen.
[Storch *et al.* (2013)]

Interkalierende Substanzen (z.B. einige Sekundärstoffe von Pfanzen zur Mikroben- und Fressfeindabwehr) sind planar und hydrophob. Lagern sie sich zwischen Basenstapel der DNA-Helix, führen sie bei Replikation zu 1bp langen Deletionen oder Insertionen. Interkalation können Einzell- und Doppelstrangbrüche verursachen.
[Storch *et al.* (2013)]

Helixformveränderungen: Die Helixform der DNA kann sich unter bestimmten Bedingungen verändern. Dann paart Adenin mit Cytosin und Tymidin mit Guanin. Das führt zu Transitionen.
[Storch *et al.* (2013)]

B.1.4.2. Selektion

Weil Mutationen Information eines lebensfähigen Organismus zerstören, sind sie viel häufiger mit negativen Konsequenzen verbunden als mit positiven. Mutationen mit negativen Folgen werden seltener beobachtet, weil die betroffenen Positionen durch Selektionsdruck konserviert werden – d.h. die betroffenen Lebewesen hätten es schwerer sich zu vermehren als ihre Verwandten ohne Mutation. Viele Mutationen sind daher in Sequenzbereichen zu finden, die wenig Auswirkung haben. Dazu gehören insbesondere viele Bereiche, die nicht für Proteine oder RNA kodieren.
[Knippers (1997)]

B.1.4.3. Prophagen:

In Bakterien besteht allerdings ein Selektionsdruck nicht-informative Chromosom-Bereiche loszuwerden. Das schränkt die Möglichkeit für nicht-negative Mutationen ein. Andererseits existieren in vielen Bakterien Mechanismen zur integration fremder DNA. Insbesondere können Phagen sich als Prophage in das Wirtsgenom integrieren und sich Generationen später in Stresssituationen wieder aktivieren. Da das Bakterium ohne Prophage optimiert war, eignet sich der erworbene Chromosom-Anteil zur Ansammlung neuer Mutationen - insbesondere Deletionen. Bakterien profitieren auch von Mutationen, die die Wiederaktivierung verhindern. Prophagen sind die bedeutensten Elemente für die Evolution mikrobieller Chromosome. Prophagen enthalten oft auch Funktionalität, die eine Infektion der Bakterie mit gleichartigen Phagen erschwert und dem Wirt auf diese Weise einen Selektionsvorteil verschafft.
[Ramisetty & Sudhakari (2019)]

B.1.4.4. Protein-kodierende Gene

Synonyme Mutationen: Auch in Protein-kodierenden Genen können geeignete Punktmutationen auftreten ohne eine Wirkung zu haben, wenn sich ein Codon verändert aber das neue Codon für dieselbe Aminosäure kodiert wie das alte [Knippers (1997)].

Bemerkung 46:
Solche Mutationen werden "synonyme Mutationen" genannt. Auf die Expressionsgeschwindigkeit können auch sie eine Wirkung haben, die der Selektion unterliegt [Graw (2015)]. Das wird meistens vernachlässigt.

Nicht-synonyme Mutationen: Punktmutationen können zu vollständigem Funktionsverlust führen, wenn beispielsweise ein Stopcodon entsteht oder eine Aminosäure im reaktiven Zentrum eines Enzyms ausgetauscht wird. ((Das reaktive Zentrum, ist der Ort, an dem das Enzym die chemische Reaktion katalysiert.)) Die Funktion eines Proteins könnte aber auch bloß leicht eingeschränkt werden. Oder die Stabilität des Proteins könnte bei höheren

Temperaturen verloren gehen, so dass nur in bestimmten Lebensräumen Nachteile entstehen.
[Knippers (1997)]

Proteinstruktur: Die Struktur eines Proteins ist stärker konserviert als die Aminosäuresequenz des Proteins. Sie basiert hauptsächlich auf Aminosäuren im Inneren des Proteins. Wenn eine Aminosäure dort verändert wird, liegt häufig mindestens eine weitere Mutation vor, die die physikalischen Änderungen durch die erste ausgleicht, so dass die Struktur erhalten bleiben kann. Aufgrund der sehr begrenzten Anzahl tatsächlich vorkommender grundlegend unterschiedlicher Strukturen, wird oft vermutet, dass neue Proteine durch Aneinandersetzen zweier bestehender Gensequenzen oder einer Verdopplung entstehen und durch weitere Mutation nur angepasst werden.
[Schulz & Schrimer (1979)]

Häufigkeiten von Punktmutationen: Aus dem genetischen Code lässt sich ableiten, dass ein Viertel der möglichen Mutationen in Protein-kodierenden Genen synomyme Mutationen wären ((wenn alle Nukleotidkombinationen und Übergänge zwischen ihnen gleich wahrscheinlich wären. Nukleotide sind nicht gleichhäufig und die Übergangswahrscheinlichkeiten sind nicht gleich – siehe Abschnitt B.1.2.2)). Transitionen sind eher synonym als Transversionen. Die weiteren Zahlenangaben für Mutationshäufigkeiten in diesem Abschnitt sind sehr grobe Schätzungen, die anhand weniger Studien durchgeführt wurden. Mutationen sind häufiger zwischen ähnlichen (gemäß Größe, Form, Flexibiltät, Ladung und Wasserstoffbrückenbindungsbildungstendenz) Aminosäuren. Mutationen wurden am häufigsten für die Aminosäure Serin beobachtet. Wenn eine Mutation in 50 Aminosäuren vorlag, war es in 4% der Fälle ein Serin. Den niedrigsten Anteil zeigte Typtophan mit 1% der Mutationen – also um den Faktor vier niedriger.
[Schulz & Schrimer (1979)]

Die Anzahl der unterschiedlichen Aminosäuren in Säuger-Genomen korreliert gut mit dem stationären Zuständ eines Markovprozesses mit gleichwahrscheinlicher Mutation zwischen Nukleotiden – mit der Ausnahme von der Aminosäure Argenin ($\rho = 0.69$, ohne Argenin: $\rho = 0.9$), die zwei bis drei mal so häufig zu erwarten wäre [King & Jukes (1969)].

Zeiträume mit weniger strikten Umweltbedingungen können zu mehr Mutationen führen als solche mit starker Selektion. Die Häufigkeit beobachteter Mutationen hängt auch von der Funktion und der Struktur DNA ab. Grobe Schätzungen für die Wahrscheinlichkeiten akzeptierter Mutationen bei Säugetieren in einem Codon in einer Milliarden Jahre sind in Tabelle B.2 angegeben. Der Häufigkeitsunterschied zwischen mRNA für Hämoglobin und hypervariablen DNA entspricht ungefähr dem, was Aufgrund synonymer Mutationen zu erwarten ist. Die zugehörige Aminosäuresequenz ist ungefähr eine Größenordnung stärker konserviert. Histon 4 ist ein extrem konserviertes Protein, das bei Eukaryoten für wesentliche Funktionen und Strukturen von Chromosomen mitverantwortlich ist. Das Protein ist

Tabelle B.2.: Anzahl akzeptierter Mutationen* aus Schulz & Schrimer (1979)

Region	Anzahl*
Hämoglobin (mRNA)	10
Hämoglobin (Aminosäurensequenz)	1.4
mRNA von Histon 4 in Seeigeln	3.5
Histon 4 (Aminosäurensequenz)	0.009
Hypervariable Proteinregionen	8.5
Hypervariable DNA-Region	50
Durchschnittliche[1] nicht-repetitive DNA	4.5 bis 8.5

Die letzten beiden beschränken sich nicht auf kodierende Bereiche. *: Grobe Schätzungen für die Wahrscheinlichkeiten akzeptierter Mutationen in einem Codon (3 Nukleotide oder eine Aminosäure) in einer Milliarden Jahren. [1]: Durchschnitt von Säuger-DNA.

nochmals zwei Größenordnungen stärker konserviert als Hämoglobin und seine Nukleotidsequenz ungefähr um den Faktor drei. In Seeigeln liegt das Gen für Histon 4 in ca. 1200 Kopien pro Zelle vor. Innerhalb einer Art ist die Sequenz sehr homogen. Zwischen Arten gibt es Sequenzunterschiede.
[Schulz & Schrimer (1979)]

Rekombination können längere Sequenzen konservieren, wenn diese mehrfach im Genom vorkommen [Hastings *et al.* (2009)].

Menschliche Spermien im selben Mann unterscheiden sich im Mittel durch 20 SNPs, während diploide Genome (6.6 Milliarden bp) zweier Menschen in geschätzten 5 Millionen Basen voneinander abweichen [Storch *et al.* (2013)].

Korrigierte aber eventuell nicht akzeptierte Häufigkeiten: Beim Menschen tritt im Durchschnitt nach 100000-200000 Genreplikationen eine Mutation auf. Menschen besitzen ca. 2 mal 20000 Gene. In einer menschlichen Keimbahmzelle treten 10-20 Basensubstitutionen pro Jahr bei 3 Milliarden bp auf. In Bakterien treten pro Generation und Nukleotid geschätzt 10^{-10} bis 10^{-9} Substitutions-Mutationen auf. Man schätzt (sowohl für Bakterien als auch Eukaryoten), dass 10^{-6} bis 10^{-5} Mutationen pro Generation in einem Gen zum Ausfall des Gens führen.
[Storch *et al.* (2013)]

Leserasterverschiebung: Insertionen, Deletionen und Translokationen können in Proteinkodierenden Genen zu Leserasterverschiebungen bei der Proteinsynthese führen. Wenn das Leseraster dabei nicht um ein Vielfaches von drei Basen verschoben wird, verändert sich die Aminosäurensequenz des betroffenen Proteins normalerweise extrem, so dass es seine Funktion vollständig verliert.
[Knippers (1997)]

B.1.4.5. Indels

Die längenverändernden Mutationen Insertion und Deletion werden auch unter dem Namen "Indel" zusammen gefasst. Es wird angenommen, dass Indels seltener auftreten als Nukleotidsubstitionen. Insbesondere gilt dies für längere Indels, die die Sekundärstruktur stärker beeinflussen. Daher könnten sie Informationen enthalten, die besonders für Analysen auf den höheren taxonomischen Ebenen nützlich ist. Es ist jedoch nicht immer verlässlich einschätzbar, ob ein zusätzlicher Sequenzabschnitt durch ein oder durch mehrere Indels entstanden ist. Das Entfernen Indel-reicher Regionen verbessert das Ergebnis einfacher Alignmentalgorithmen. Wenn Indels allerdings geeignet berücksichtigt werden, können sie die phylogenetische Auflösung verbessern. Das ist jedoch aufwendig.
[Nagy *et al.* (2012)]

B.1.4.6. Kopienzahl Variationen

Kopienzahl-Variationen (CNV) sind Unterschiede (zwischen Genomen) in der Häufigkeit, mit der eine Sequenz im chromosomalen Genom vorkommt. Die Sequenzen brauchen dabei nicht vollständig identisch zu sein und enthalten häufig Indels. ((Der Begriff CNV wird auch für Mutationen gebraucht, die zu CNVs führen.)) CNVs verändern die Struktur von Chromosomen. Etwa 12% des menschlichen Genoms unterliegt CNV. CNV zeigt sich als ein Hauptfaktor der Evolution – insbesondere bei schnell-ablaufender Evolution. CNV ändern die Anzahl der Expressionen von enthaltenen Genen. Außerdem entsteht durch zusätzliche Kopien eine Redundanz. Einige Kopien können neue oder veränderte Funktionen entwickeln. Sie können auch andere Expressionsmuster zeigen, also in anderen Situationen oder zusammen mit anderen Genen exprimiert werden. Wenn eine CNV durch ein nichthomologes Rekombinationsereignis entsteht, können auch Sequenzen zwischen verschiedenen Genen kopiert werden, die Proteinen weitere Funktionen oder Bindungsstellen geben können. Oft wird CNV jedoch nur schlecht toleriert. Trotzdem ist die Wahrscheinlichkeit, dass sich die Kopienzahl an einem von CNV betroffenen gewählten Lokus ändert, mehrere Größenordnungen höher als die für eine Punktmutation an einer betrachteten Base. Die zugrundeliegenden Mechanismen der Entstehung unterscheiden sich nicht wesentlich zwischen Menschen, Bakterien und Hefepilzen. Abgesehen davon, dass überproportional viele CNVs bei der geschlechtlichen Teilung entstehen ((die nicht in Bakterien und nicht allen Hefen existiert)). Bakterien enthalten allerdings praktisch keine repetitiven DNA-Elemente [Storch *et al.* (2013)].

CNVs werden in zwei Gruppen eingeteilt:

Recurrent CNVs entstehen durch (nicht-allelische) homologe Recombination (siehe Abschnitt B.1.2.6) zwischen (direkt) wiederholten Sequenzen. Sie zeigen eher geringe Kopienzahlen. Sie sind wenige Kilobasenpaare lang und weisen gewöhnlich über 95% Übereinstimmung auf. CNVs können insbesondere dann entstehen, wenn während der DNA-Replikation die beteiligten Molekülkomplexe kaputtgehen. Unter Stress kann dann zu einem ungenauen nicht-homologen Reparaturprozess gewechselt werden der CNVs begünstigt.

Non-recurrent CNVs entstehen durch nicht-homologe Mechanismen, die über das gesamte Genom hinweg wirken können. Sie zeigen oft nur geringe Homologie (2 bis 15 bp) an ihren Endpunkten und weisen eine komplexe Struktur auf – häufig viele Indels, Inversionen, Wiederholungen und Stücke aus anderen DNA-Bereichen.

In Menschen liegen CNVs meist mit wenigen Wiederholungen (die auch invertiert sein können) geclustert in Regionen mit komplexer genomischer Architektur.
[Hastings *et al.* (2009)]

Einflussfaktoren der DNA-Struktur: NAHR-vermittelte Änderungen geschehen an Stellen, die bereits eine Wiederholungen aufweisen, da sie auf Homologie angewiesen sind. CNVs treten häufiger im Heterochromatin sowie Strukturen wie Replikationsursprüngen, Terminatoren und Sequenzen auf, die die Anlagerung von Molekülen begünstigen.
[Hastings *et al.* (2009)]

B.1.5. Co-Evolution

Mutationen können gelegentlich zu besseren Anpassungen an die Umwelt führen und somit zu einem evolutionären Vorteil. Sich ändernde Umwelteinflüsse stehen der Entwicklung eines perfekt-angepasstem Organismus entgegen. Ein besonders genetisches Beispiel ist die Co-Evolution zwischen Parasit und Wirt – ein Wettrüsten. Dabei kann sich Schädlichkeit des Parasiten erhöhen oder vermindern, wenn es seine Fitness erhöht.
[Begon *et al.* (2016)]

Bakterien haben viele Methoden zur Abwehr von Phagen entwickelt. Davon sei hier eine erwähnt, die die DNA direkt betrifft und erwähnte Konzepte aufgreift und verdeutlicht.

Phagen, Bakterien, Restriktion und lateraler Gentransfer: So hat sich in Bakterien ein ausgeprägtes System von DNA-schneidenden Enzymen (sog. "Restriktionsenzyme") entwickelt (und entwickelt sich noch weiter), das die DNA von Bakteriophagen zerschneidet und unschädlich macht. Bakteriophagen sind nun evolutionären Druck ausgesetzt ihre DNA so aufzubauen, dass sie keine Erkennungssequenzen für diese Enzyme aufweist und nicht mehr zerschnitten wird. Weshalb wiederum Restriktionsenzyme angepasst werden müssen. Es ergibt sich ein Kreislauf. Wenn der Wirt das Wettrüsten gewinnt, ist das Überleben der Phagenpopulation gefährdet. Einige Bakteriophagen haben Methoden entwickelt, mit denen sie gezielt die Mutationsrate in einem Bereich ihrer DNA erhöhen, der für die Wirts-Erkennung verantwortlich ist. Auf diese Weise können sie (als Population) zwischen verschiedenen Wirten wechseln. Die Restriktionsenzyme der Bakterien sind andererseits auch für transponierende Elemente nötig, die sowohl eine Hauptursache für Mutationen darstellen, als auch den Austausch von DNA zwischen Bakterien wesentlich begünstigen. Insbesondere werden auch Resistenz-Gene gegen Bakteriophagen über Plasmide ausgetauscht. Bakteriophagen kommen häufiger vor als Bakterien und werden schneller erzeugt. Diesen

Vorteil können Bakterien durch die Übermittlung von DNA (lateraler Gentransfer) untereinander abschwächen.
[Labrie *et al.* (2010)]
Außerdem haben Prokaryoten (siehe Abschnitt B.1.6) ein CRISPR genanntes Abwehrsystem entwickelt. Dabei handelt es sich um einen Bereich ihres Chromosoms, in den kurze DNA-Fragmente (23-55 bp) von Phagen eingebaut werden. Aus diesen werden RNAs entwickelt, die sich an die passende Phagen anlagern und diese unschädlich machen können.
[Graw (2015)]

Bemerkung 47:
Insbesondere CRISPR enthält also Informationen über die Phagen, mit denen ein Bakterium oder seine Vorfahren in Kontakt gekommen sind.

Als Antwort auf Restriktionsenzyme hat zumindest ein Bakteriophage eine abgeänderte Form von DNA entwickelt. Anstelle von Cytosin (C) ist in seiner DNA die Base Hydroxymethylcytosin zu finden. Einige Bakterien haben Restriktionsenzyme entwickelt, die Hydroxymethylcytosin-enthaltende DNA schneiden können.
[Labrie *et al.* (2010)]

Bemerkung 48:
Nicht jedes Sequenzierverfahren ist standardmäßig geeignet diese Base zu erkennen.

B.1.6. Pro- und Eukaryoten

Die größten Unterschiede im genetischen Prozessablauf definieren die grundlegende Einteilung von Lebewesen. Zunächst wird zwischen solchen mit Zellkern (Eukaryoten) und ohne Zellkern (Prokaryoten) unterschieden wobei sich letzte in zwei sehr unterschiedliche Gruppen aufspalten, so dass insgesamt von den drei Domänen des Lebens gesprochen wird.
[Graw (2015)]

B.1.6.1. Prokaryoten

Prokaryoten werden aufgeteilt in Bakterien und Archaeen. Bakterien sind zahlreicher und für viele menschliche Interessen von größerer Bedeutung. Fast alle Prokaryoten bestehen aus einer einzelnen Zelle, jedoch gibt es Bakterien (Myxobakterien), die zeitweise mehrzellig leben [Storch *et al.* (2013)]. Es wurde angenommen, dass Prokaryoten genau ein Chromosom besitzen, das als geschlossener Ring eines Doppelstrangs vorliegt.
[Raven *et al.* (2000)]
Es wurden aber Bakterienarten entdeckt, die zwei Chromosome haben [Heidelberg *et al.* (2000)]. Das Bakterium *Agrobacterium tumefaciens* enthält sowohl ein ringförmiges als auch ein lineares Chromosom. Die Größe bekannter bakterieller Chromosome liegt zwischen $5.8e5$ und $9.1e6$ bp. Das Chromosom des Bakteriums *E. coli* ist $4.6e6$ bp lang (das entspricht ungefähr einem Millimeter) und enthält 4381 Gene (4288 kodieren für Proteine, 7 für rRNA und 86

für tRNA), die durch etwa 100 bp lange nicht-kodierende Sequenzen getrennt sind. Die Richtungen, aus der sie abgelesen werden müssen, unterscheidet sich zwischen Genen. Neben dem Chromosom können in Prokaryoten auch kleinere DNA-Ringe (Plasmide) vorliegen.
[Graw (2015)]

Plasmide können DNA-Fragmente mit dem Chromosom austauschen und auf andere Prokaryoten übertragen werden. Aufgrund ihres einfachen Aufbaus können sich Prokaryoten sehr schnell vermehren. Ein Rekreationszyklus kann beim Bakterium *E. coli*, unter optimalen Bedingungen, innerhalb von ca. 20 Minuten abgeschlossen sein. Dies ermöglicht eine schnelle Evolution, Selektion und Resistenz der Population gegenüber hohen Mutationshäufigkeiten pro Zeit.
[Knippers (1997)]
Viele, aber nicht alle, Bakterien können reine DNA aus ihrer Umweld aufnehmen und in ihr Genom einbauen. Dieser Vorgang wird Transformation genannt. Voraussetzung für die Übernahme der DNA auf diese Weise ist unter anderem eine teilweise Sequenzübereinstimmung der Spender- und Wirts-DNA, da der Einbau durch homologe Rekombination geschieht.
[Wehner & Gehring (1993)]

B.1.6.2. Eukaryoten

Eukaryoten werden aufgeteilt in: Protisten ("Einzeller"), Pilze (ein- oder mehrzellig), Pflanzen und Tiere. ((Nicht jedes Lebewesen, das nur aus einer Zelle besteht, ist ein Einzeller.)) Eukaryotische Zellen enthalten in ihrem inneren Strukturen, die von Membranen umschlossen sind. Insbesondere gilt dies für den Zellkern. Der Zellkern enthält mehrere Chromosomen. Die Chromosomen können in mehrfacher Ausführung vorliegen. Bei normalen menschlichen Frauen liegen beispielsweise alle Chromosomen des Zellkerns in doppelter Ausführung vor – jeweils eins vererbt durch den Vater und eins durch die Mutter. Dies hat unter anderem den Vorteil, dass, wenn ein Gen aufgrund von Mutationen funktionsunfähig wird, dasselbe Gen noch auf einem weiteren Chromosom existiert. Bei sexueller Vermehrung finden homologe Rekombinationsereignisse zwischen gleichartigen Chromosomen statt. Auf diese Weise werden DNA-Fragmente zwischen den Chromosomen ausgetauscht. In viele Pilzzellen liegen die Chromosome nur einfach vor [Dutheil (2020)]. In einigen Eukaryoten (z.B. Amphibien) können, abhängig von der Lebensphase, auch mehr als eine Millionen Kopien von chromosomalen (rRNA-)Genen in kleinen geschlossen Ringen angefertigt werden. Diese Kopien werden nicht vererbt. Einige Eukaryoten (z.B. der Pilz *Saccharomyces cervisiae*) besitzen echte Plasmide. In einigen weiteren membranumschlossen Strukturen (z.B. Mitochondrien), innerhalb der Zelle aber außerhalb des Zellkerns, existiert ein eigenes Chromosom.
[Lewin (2000), Raven *et al.* (2000), Campbell *et al.* (1999), Wehner & Gehring (1993)]

Transkription in Eukaryoten: Transkription ist in Eukaryoten komplizierter als in Prokaryoten. Bei der Transkription wird zunächst prä-mRNA erzeugt. Diese enthält Abschnitte, die für das Protein kodieren (Exons) und Abschnitte, die dies nicht tun (Introns). Die Introns werden aus der prä-mRNA herausgeschnitten und die Exons zusammengesetzt. Dieser

Prozess wird Splicing genannt. Manchmal wird dieselbe prä-mRNA zu unterschiedlichen mRNAs zusammengesetzt (alternatives Splicing) und kann hierdurch für verschiedene Proteine oder Proteinuntereinheiten kodieren.
[Lewin (2000)]

Bemerkung 49:
Auch die mitochondriale DNA (also die aus den Mitochondrien) könnte theoretisch für Mikrobiomanalysen, z.B. bei Pilzen, genutzt werden, allerdings existieren sehr viele Mitochondrien in einer Zelle.

B.1.7. Ribosomen

Ribosomen bestehen hauptsächlich aus ribosomaler RNA (rRNA) und setzen sich aus einer großen und einer kleinen Untereinheit zusammen. Die große Untereinheit besteht wiederum aus weiteren Untereinheiten, die kleine Untereinheit nicht. Zentrifugation der kleinen Untereinheit ergibt bei Eukaryoten einen Wert von 18S und bei Prokaryoten 16S. "S" ist hierbei ein Maß für die Auftrennungsposition nach Zentrifugation und kann näherungsweise so interpretiert werden, dass eine kleinere Zahl einem kleineren Objekt entspricht. Trotz derselben Position nach Zentrifugation unterscheiden sich auch die 16S-Untereinheiten von Bakterien und Archaea. DNA-Regionen, die für Ribosomen kodieren, sind weitgehend stark konserviert. Ribosomen spielen eine zentrale Rolle bei der lebenswichtigen Translation. Bei der Translation ist das Zusammenspiel und räumliche Zusammenpassen insbesondere von Ribosom, mRNA und tRNA notwendig. Da Mutationen dies gefährden, besteht ein hoher Selektionsdruck zur Beibehaltung der Sequenz. Insbesondere gilt dies für die Bestandteile des Ribosoms, die aus rRNA bestehen. Im Gegensatz zur Proteinsynthese unterläuft die Synthese der rRNA keiner Translation. Jede Änderung der rRNA-kodierenden DNA-Sequenz hat daher eine Auswirkung auf das Produkt. Es sei denn, der entsprechende Sequenzabschnitt wird vorher entfernt.
[Knippers (1997), Jay & Inskeep (2015)]

B.1.7.1. rRNA

Spacer in Bakterien: Anzahl und Aufbau von rRNA Genen unterscheidet sich zwischen Organismen. Das Bakterium *E. coli* enthält sieben rRNA-Gene. Die Gene sind sich ähnlich aber nicht identisch. Jedes enthält einen Abschnitt für 16S rRNA, 23S rRNA und 5S rRNA. Diese sind durch kurze Sequenzen voneinander getrennt, die "Spacer" genannt werden. Im Spacer zwischen 16S und 23S liegen, unter anderem, zwei Gene für tRNAs. Das gesamte rRNA Gen wird komplett als ein Stück transkribiert und im Anschluss in seine Bestandteile zerlegt.
[Knippers (1997)]

Schleifen-bildende RNA-Sequenzen, die auch in 16S- und 18S-Regionen vorkommen, enthalten vergleichsweise viele SNPs, Indels und Inversionen, die Alignments erschweren [Storch *et al.* (2013)].

Internal Transcribed Spacer: In Eukaryoten kommen ebenfalls Spacer vor – hier zwischen den Sequenzen für die 18S, 28S und 5.8S ribosomalen Untereinheiten. Im Gegensatz zu RNA-kodierenden Sequenzabschnitten unterscheiden sich diese Spacer in ihrer Länge deutlich zwischen verschieden Eukaryoten. Die Längenunterschiede kommen durch unterschiedliche Anzahlen identischer, sich wiederholender Sequenzen zustande. Diese Spacer werden "Internal Transcribed Spacer" (ITS) gennant. Das 5.8S Gen wird zusammen mit den beiden es umgebenen Spacern als ITS-Region bezeichnet. Diese wird unter anderem bei Mikrobiomstudien für Pilze verwendet.
[Schoch *et al.* (2012), Schreiter *et al.* (2014)]

Bedeutung für Phylogenetik: Die Eigenschaften von Ribosomen machen rRNA-Gene zu attraktiven Zielen für phylogenetische Untersuchungen. Die 16S-Region wurde insbesondere für die Definition der drei Domänen des Lebens (Bakterien, Archaeen und Eukaryoten) verwendet ((um Bakterien von Archaeen zu unterscheiden)). Für die Analysen zu Archaeen werden andere 16S Primer benutzt als für Bakterien.
[Jay & Inskeep (2015)]

B.1.8. Lateraler Gentransfer

Anpassungen an ökologische Bedingungen und kompetitives Aufrüsten erfordern viel Innovation. Lateraler Gentransfer kann als schnelle Evolution betrachtet werden, denn Organismen können sich auf diese Weise schnell an geänderte Umweltbedingungen anpassen. Die drei häufigsten Mechanismen für lateralen Gentransfer sind: Transformation, Transduktion und Konjugation. Transformation bezeichnet die Aufnahme von DNA direkt aus der Umwelt, Transduktion, die Übertragung durch Viren und Konjugation, die direkte Übertragung zwischen zwei Zellen. Transduktion kann auch zwischen Organismen geschehen, die in unterschiedlichen Umgebungen leben.
[Medini *et al.* (2005)]
In Prokaryoten stammen bis zu 32% der Gene aus noch nicht allzulange zurückliegendem lateralem Gentransfer und über 75% prokaryotischer Gene sind von lateralem Gentransfer betroffen. Erhöhte Raten von lateralem Gentransfer stehen bei Eukaryoten in Zusammenhang mit dem Verlust von sexueller Reproduktion und dem von aerobem Stoffwechsel ((d.h. Sauerstoffverträglichkeit)). Der Aufbau eukaryotischer Zellen - z.B. der Zellkern - stellt ein Hindernis für lateralen Gentransfer dar. Lateraler Gentransfer tritt viel seltener auf als Genverlust oder -duplikation. Er hat aber großen Einfluss auf die Artbildung von Bakterien und ihre Phylogenie [Storch *et al.* (2013)]. Das Aufspüren von lateralem Gentransfer ist schwierig und sollte einen Vergleich von phylogenetischen Bäumen enthalten, die einmal davon ausgehen, dass er vorlag und einmal davon, dass er nicht vorlag. Dabei ist auch darauf zu achten, ob eine betroffene Sequenz stärker konserviert ist, als es von der angenommenen Zeitdauer her seit der Abspaltung in dem phylogenetischen Baum zu erwarten wäre.
[Wisecaver & Rokas (2015), Rokas *et al.* (2018)]

Lateraler Gentransfer geschieht auch zwischen sehr unterschiedlichen Organismen. Die

Pflanzenfamilie der Hülsenfrüchtler besitzt Gene für Hämoglobin. Ca. 100 Gene im Mensch werden direkt auf Bakterien zurückgeführt. Auch Gentransfer von Mensch zu Bakterium findet statt.
[Storch *et al.* (2013)]

B.1.8.1. Lateraler Gentransfer in Pilzen

Lateraler Gentransfer findet in Pilzen, Pfanzen und Tieren statt, bei weitem am häufigsten jedoch in Prokaryoten. Pilze erhalten lateralen Gentransfer nicht nur von Bakterien sondern auch von Pflanzen, Protisten und (wahrscheinlich am Häufigsten) von Pilzen. Es wird geschätzt, dass zwischen 0.1 und 2.8% der Gene in einem typischen Pilz von lateralem Gentransfer betroffen sind, wobei diese Schätzung zu niedrig sein könnte, da Pilzgenome noch nicht gut verstanden werden. Einige Pilze haben ganze Chromosome, die nicht lebensnotwendig sind, und vermutlich aus lateralem Gentransfer stammen.
[Wisecaver & Rokas (2015), Rokas *et al.* (2018)]

B.1.9. Organisation durch natürliche Selektion

B.1.9.1. Metabolische Gencluster

Produkte eukaryotischer Gene stehen häufig in komplizierten Netzwerkbeziehungen. Lateraler Gentransfer tritt seltener bei Genen auf, wenn deren Produkte an komplexen Zusammenhängen beteiligt sind. Das Zusammenbringen solcher Gene in einen Sequenzabschnitt (dann ein "Gen-Cluster") ermöglicht lateralen Gentransfer, bei dem die Zusammenhänge erhalten bleiben. Auf diese Weise wird die Häufigkeit an lateralem Gentransfer der betroffenen Gene anscheinend auf das ca. 1.7-fache erhöht. Insbesondere sogenannte metabolische Gene treten oft in Clustern auf. Diejenigen, die an der Schädigung anderer Organismen beteiligt sind, werden häufiger lateral übertragen.
[Wisecaver & Rokas (2015), Rokas *et al.* (2018)]

Entstehung und Verlust: Populationsgröße, -struktur und Paarungssystem beeinflussen über natürlicher Selektion die Struktur eukaryotischer Genome, einschließlich der Existenz von metabolischen Genclustern. Die Entstehung von Genclustern erfordert häufig die Umstrukturierung großer chromosomaler Segmente. Metabolische Gencluster liegen oft in sich schnell entwickelnden Chromosomabschnitten, z.B. in der Nähe von Telomeren oder mobilen genetischen Elementen.

Da Cluster metabolischer Gene meist nicht lebensnotwendig sind, sondern nur einen Vorteil unter bestimmten Bedingungen bringen, können sie durch Selektion auch verloren gehen. Pathogene eukaryotische Mikroorganismen neigen dazu ein reduziertes Genom mit eingeschränktem Metabolismus zu entwickeln und auf enge ökologische Bedingungen fixiert zu

sein. Es existieren drei kompatible genetische Modelle, die an der Bildung und Erhaltung der Cluster beteiligt sein könnten:

Co-Regulation: in Clustern lassen sich Gene besser steuern.

Genetische Verknüpfung: wenn Gene gut aneinander angepasst sind, profitiert der Organismus davon, dass sie gemeinsam vererbt werden. ((Die Wahrscheinlichkeit voneinander getrennt zu werden, steigt bei Eukaryoten mit zunehmenden Abstand entlang des Chromosoms.)) Außerdem zeigen Regionen, die die Cluster enthalten, vergleichsweise niedrige Rekombinationsraten.

Eigennütziger Cluster: da Cluster häufiger an erfolgreichem lateralem Gentransfer beteiligt sind, vermehren sich dadurch besser. Das Schicksal des Clusters wird von dem anderer Gene des Organismus getrennt und bekommt die Möglichkeit auch in einer anderen Art zu überleben.

[Wisecaver & Rokas (2015), Rokas *et al.* (2018)]

Metabolische Gencluster in Pilzen: Bei weitem die meisten metabolischen Gencluster befinden sich in Pilzen. Pilzgenome unterscheiden sich häufig voneinander dadurch, dass ganze Cluster vorhanden oder abwesend sind. Pilze verdauen meist andere Organismen. Wenn andere Pilze die Codierung, Herstellung und Sekretion spezifischer schädigender Proteine betreiben, hift dies auch Pilzen, die sich diesen Aufwand sparen. Wenn in metabolischen Pathways Zwischenprodukte gebildet werden, die für den Pilz selber giftig sind, liegen die Gene häufiger als Cluster vor. Wenn nur die Gene verloren gehen, die so ein Zwischenprodukt weiterverarbeiten, entsteht ein großer selektiver Nachteil. Identische metabolische Cluster sind in sehr engen taxonomischen Bereichen zu finden und können innerhalb nahverwandter Arten variieren.
[Wisecaver & Rokas (2015), Rokas *et al.* (2018)]

B.1.10. Marker-Gene

B.1.10.1. 16S als Marker

Das 16S rRNA Gen wird seit 1990 für mikrobielle Analysen sequenziert. Es ist der meist verwendete molekulare Marker in der mikrobiellen Ökologie. Für Mikrobiom-Analysen werden von dem ca. 1500 bp langen Gen nur wenige hundert Basenpaare benutzt. Evolutionär wird es konserviert und ist in allen Bakterien vorhanden. Daher kann es universell für die Identifikation von Bakterien verwendet werden, jedoch existieren auch Bakterien, die trotz identischer 16S Sequenz, sehr unterschiedlichen Genome haben.
[Case *et al.* (2007)]

Auswahl als Marker: Die Analyse von 16S, so wie auch viele andere Methoden der mikrobiellen-Ökologie, wurde ursprünglich für die Phylogenetik entwickelt. Das hat Vor- und Nachteile. Es kann auf große, bereits existierende 16S Datenbanken zurückgegriffen werden. Methoden und Abläufe sind gut etabliert und erprobt. Bausteine zur Durchführung der Analysen im Labor werden bereits kommerziell angeboten. Die 16S Sequenz ist, abhängig vom Bakterium, mehrfach an verschiedenen Stellen des Genoms vorhanden. Beobachtet wurden 1 bis 15 Positionen. Häufigere Sequenzabschnitte lassen sich im Labor leichter nachweisen als solche, die nur einmal vorkommen. Für Mikrobiomanalysen führt diese CNV offensichtlich zu Verzerrungen.
[Case *et al.* (2007)]

B.1.10.2. Proteinkodierende Marker

Das Gen für die RNA-Polymerase β Untereinheit (rpoB) ist ein Beispiel für ein proteinkodierendes Gen, das universell als Alternative für 16S eingesetzt werden kann. Es existiert nur eine bekannte Position und umgeht daher den Bias, der durch Kopien und intragenomische Heterogenität entsteht.
[Case *et al.* (2007)]

B.2. Evolution

Evolution verläuft nicht immer gleich schnell oder auf die gleiche Weise. So haben Bakterien zum Beispiel Mechanismen entwickelt die Mutationen, Rekombinationen und lateralen Gentransfer unter Stress fördern. Und Selektion wirkt je nach Situation: gerichtet, wenn sich Populationen an neue Umweltbedingungen anpassen müssen, stabilisierend, wenn die Umwelt stabil ist und Abweichungen vom verbreitetem Genotyp Nachteile bewirken, disruptiv, wenn sich Parasiten oder Räuber immer wieder auf die verbreitetsten Genotypen spezialisieren oder ausgleichend, wenn durch Genduplikation die Bedeutung des einzelnen Gens abnimmt.
[Storch *et al.* (2013)]

Bemerkung 50:
Die Evolution von Pflanzen ist besonders und wird daher hier vernachlässigt. Es wird vermutet, dass die Hälfte der höheren Pflanzen durch Hybridisierung verschiedener Arten entstanden sind. Hybriden sind zwar meist steril aber Pflanzen können sich zum einen vegetativ vermehren, indem ein Pflanzenbruchstück neu aussprosst und zum anderen vertragen Pflanzen vielfache Chromosomensätze (mehr als zwei) und können ihre Fortpflanzungsfähigkeit wiederherstellen. [Storch et al. *(2013)]*
Es wurden auch häufig schon Hybridisierungen, die zu lateralem Gentransfer führten, zwischen Pilzen festgestellt. Wenn im Chromosom ein Mosaik-artiges Muster der Gene der Genome verschiedener Arten zu sehen ist, wird auf eine solche Hybridisierung geschlossen. Die Entstehung

einiger neuer, besonders schädlicher Pflanzen-Pathogene, wird auf Hybridisierungen zurückgeführt.
[Dutheil (2020)]

Evolution kann sehr schnell wirken. In Laborversuchen zwischen Bakterien und Phagen wurden in nur ca. 100 Bakteriengenerationen mehrere Co-evolutive Wettrüstzyklen beobachtet.
[Begon *et al.* (2016)]

In kurzen Zeiträumen besteht diese Evolution hauptsächlich aus Selektion geeigneter Genotypen. Die Entwicklung geeigneter Genotypen geschieht in der Regel durch Rekombination existierender Allele. Die Entstehung positiver Mutationen ist selten. In der als besonders schnell geltenden Co-Evolution zwischen Parasit und Wirt sind meist Kombinationen von Allelen entscheidend. Zu den Ausnahmen gehört Medikamentenresistenz, die häufig auf einem einzigen Gen beruht.
[Lucius *et al.* (2018)]

B.2.1. Zeitspanne der Evolution

Die erste zellulare Lebensform ist vor über 3.9 Milliarden Jahren entstanden. Vor mehr als 3.4 Milliarden Jahren haben sich Prokaryoten in Bakterien und Archaeen differenziert. Zu diesem Zeitpunkt lebte der letzte gemeinsame Vorfahre aller heutigen Zellen. Vor weniger als 1.84 Milliarden Jahren sind Eukaryoten entstanden. Mitochondrien waren im Zeitraum von 2.053–1.21 Milliarden Jahren eigenständige, symbiotisch lebende Prokaryoten.
[Betts *et al.* (2018)]

Vor 700 Millionen Jahren entstanden Pilzen und mehrzellige Tiere. Die Zeit vor 510–570 Millionen Jahren entspricht ungefähr der Aufteilung in Stämme (Phylum) in taxonomischen Einordnungen.
[Raven *et al.* (2000)]

B.2.2. Mikrobieller Genpool

Organismen können selbstständig Gene mit neuen Funktionen entwickeln, indem ein Sequenzabschnitt verdoppelt wird und Mutationen nachfolgend zu einer Funktionsänderung einer der beiden Abschnitte führen. Allerdings ist der häufigste Weg, auf dem ein Organismus Gene mit neue Funktionen erlangt, der, dass die Gene von einem anderen Organismus übernommen werden. Der große mikrobielle Genpool lässt vermuten, dass die meisten Funktionen in Mikroben entwickelt und von höheren Organismen übernommen wurden. Dies entspricht unterschiedlichen Rollen in der Evolution.
[Medini *et al.* (2005)]

Tabelle B.3.: Grenzwerte prokaryotischer Taxonomie aus Yarza *et al.* (2014)

Taxonomische Ebene	Sequenzübereinstimmung
Stamm (= Phylum)	75.0%
Klasse	78.5%
Ordnung	82.0%
Familie	86.5%
Gattung	94.5%

B.2.3. Prokaryotische Taxonomie

Es gibt verschiedene Möglichkeiten taxonomische Einteilungen zu definieren, die zu unterschiedlichen Ergebnissen führen.

Einfach: Das 16S Gen ist momentan der einzige ausreichend unterstützte taxonomische Marker für alle Bakterien und Archaeen, der auf einer einzelnen DNA-Region beruht. Für taxonomische Einteilungen oberhalb der Artenebene von Prokaryoten wurden minimale Ähnlichkeitsgrenzwerte für die 16S Gen-Sequenz vorgeschlagen (Tab.: B.2.3). Zwei Prokaryoten, die eine niedrigere 16S Sequenzähnlichkeit als der Grenzwert für eine der taxonomischen Ebenen aufweisen, gelten für diese Ebene als verschieden. Ebenso gelten sie als verschieden, wenn andere Information dagegen spricht. Diese Information kann genetisch, phänotypisch oder ökologisch sein.
[Yarza *et al.* (2014)]

Probleme: 1. Für unvollständige Sequenzen aus Metagenom-Analysen sind die Grenzen nicht anwendbar. 2. Kleine ribosomale Untereinheiten sind von ((CNV mit intragenomischer Heterogenität)), lateralem Gentransfer, konvergenter Evolution (d.h. dasselbe Produkt entsteht unabhängig voneinander in unterschiedlichen Abstammungslinien) und unterschiedlich schneller Evolution betroffen. Einfache Sequenzvergleiche spiegeln daher manchmal nicht die korrekte Phylogenie wieder.
[Wu *et al.* (2013)]
Die Auflösung auf den höchsten und niedrigsten taxonomischen Ebenen ist gering. 3. Bei der Sequenzierung entstehen durch den PCR-Prozess oft zusammengesetzte DNA-Fragmente aus verschiedenen Vorlagen (sog. “Chimären”), die die Klassenbildung verzerren.
[Parks *et al.* (2018)]

Komplex: Aus mehreren protein-kodierenden Genen, die so ausgewählt wurden, dass sie wenig Probleme verursachen, wird eine künstliche Sequenz aneinandergereiht und ein Alignment angefertigt. Aus diesem wird ein phylogenetischer Baum konstruiert. Es werden Evolutionsraten (pro Gen und Organismus) geschätzt und berücksichtigt. Die taxonomischen Ebenen werden gleichmäßig zugeordnet oder es wird sich an Datenbanken mit bekannten

Einträgen orientiert.
[Wu *et al.* (2013)]
Bestehende Datenbanken wurden auf diese Weise auch verbessert [Parks *et al.* (2018)]. Wenn phylogenetische Untersuchungen auf rDNA Alignments basieren, wird eine manuelle Korrektur vorgenommen, die die räumliche Struktur der rRNA berücksichtigt [Storch *et al.* (2013)].

Datenbank-basiert: Die grundlegende taxonomische Zuordnungen erfolgt in Einzelfallentscheidungen durch Experten und wird der Öffentlichkeit in Datenbanken zur Verfügung gestellt. Diese Methode findet auch bei Eukaryoten Anwendung.
[Federhen (2011)]
Daten aus Experimenten werden anhand ihrer Sequenzen gegen diese Datenbanken verglichen und die taxonomische Zuordnung aus diesen übertragen [Antweiler *et al.* (2017)].

B.2.3.1. Stammklassifikation

Bakterien lassen sich auch unterhalb der Arten-Ebene auf verschiedene Weisen klassifizieren. Eine pragmatische Einteilung ist die "Serogruppe", die auf preiswerten immunochemischen Verfahren beruht. Hierbei wird nur die Zusammensetzung der Bakterienoberfläche berücksichtigt. ((Diese Einteilung ist nicht genetisch sondern phänotypisch.)) Für eine genetischen Einteilung wird oft das Verfahren "Multi-Locus Sequence Typing" (MLST) verwendet, bei dem Unterschiede in Housekeeping-Genen bestimmt werden. Das Ergebnis ist der "Sequenztyp" des Bakteriums. Sequenzvergleiche über das gesamte Genom haben ergeben, dass diese Einteilungsmethoden nicht der genomischen Diversität entsprechen. Häufig sind Bakterien verschiedener Serogruppen einander ähnlicher als solche innerhalb einer Serogruppe. Auch Genome desselben Sequenztyps können sich sehr von einander unterscheiden. Die Klassifikation von Bakterien ist teilweise mehr an der medizinischen Bedeutung als an Verwandtschaftsbeziehungen orientiert. Zum Beispiel unterscheidet sich *B. anthracis* von *B. cereus* hauptsächlich durch zwei Plasmide. Eins dieser Plasmide kodiert für Anthrax und ist daher von medizinischer Bedeutung. Zwei Bakterien aufgrund von Plasmiden in unterschiedliche Arten (oder auch nur Stämmen) aufzuteilen, widerspricht dem Konzept genetischer Definitionen.
[Medini *et al.* (2005)]

***E. coli*:** Das Kerngenom von *E. coli* umfasst ca. 2000 Gene. Das Pan-Genom ca. 18000. Die meisten Vertreter besitzen zwischen 4200 und 5500 Genen. Die beiden Stämme K12 und EHEC (enteroaggregatives *E. coli*) haben nur 70% gemeinsame Gene. Von ihrer medizinischen Bedeutung unterscheiden sich die Stämme ebenfalls. Der EHEC-Erreger des Ausbruchs von 2011 in Deutschland wurde aus medizinischer Sicht als EHEC klassifiziert. Er war eher ein EAEC (enteroaggregatives *E. coli*) das ein Toxin-Gen von EHEC, mehrere Gene des extratestinalen pathogenen E. coli (ExPEC) und mehrere Antibiotika-Resistenzgene erworben hatte.

Tabelle B.4.: Grenzwerte fungaler Taxonomie aus Tiew *et al.* (2020)

Taxonomische Ebene	Sequenzübereinstimmung
Klasse	80.9%
Ordnung	81.2%
Familie	88.5%
Gattung	94.3%

[Storch *et al.* (2013)]

Bemerkung 51:
Die Klassifikation anhand Übereinstimmungsquoten konservierter DNA-Regionen, kann erhebliche Variabilität außerhalb dieser Regionen zulassen und Organismen zusammenfassen, die sehr unterschiedlich auf ihre Umgebung wirken.

B.2.4. Pilz-Taxonomie

Pilztaxonomie ist noch nicht gut entwickelt. Vergleichsdatenbanken sind dünn besetzt. Dies führt beim Next Generation Sequencing zu vielen unklassifizierten OTUs. Korrekte Identifikation ist auf Art-Ebene mit über die Sequenz der ITS-Region zur Zeit nicht möglich. Sequenziermethoden der dritten Generation können die Identifikation von Pilzen verbessern, da längere Sequenzen ausgelesen werden. ITS-Primer zeigen Primer-Bias, Amplifikations-Bias und können Mock-Gemeinschaften nicht korrekt wiedergeben.

Für taxonomische Einteilungen von filamentösen Pilzen wurden Grenzwerte für die ITS-Region-Sequenz geschätzt (Tab.: B.2.4).
[Tiew *et al.* (2020)]

B.2.5. Entstehung mikrobieller Arten

Das periodische Selektionsmodell ist das klassische Modell mikrobieller Evolution. Ihm zufolge entstehen Mutationen mit großem positiven Einfluss in asexuellen Populationen nur selten. Das Individuum mit dieser Mutation setzt sich gegenüber anderen durch und sein Genotyp verdrängt alle anderen, wenn keine oder kaum Rekombination stattfindet. Hierdurch geht der Population jede Diversität verloren. Alle beobachtete Diversität innerhalb einer Population muss daher nach ((und während)) der Fixierung des letzten großen Effekts entstanden sein. Das Modell nimmt an, dass positiv-wirkende Mutationen so selten sind, dass jeweils höchstens eine vorliegt. Experimente lassen hingegen vermuten, dass mehrere positiv-wirkende Mutationen gleichzeitig in Individuen auftreten.
[Wilmes *et al.* (2008)]

Das klonale Ökotypmodell ist eine Erweiterung des periodische Selektionsmodells. Es nimmt an, dass die seltene positive Mutation die Fitness für eine bestimmte Nische in einem Ökosystem erhöht und der mutierte Genotyp in dieser Nische fixiert wird, in anderen aber nicht. Gruppen von Individuen mit einheitlichen Sequenzen innerhalb einer Art entsprechen demnach hauptsächlich Populationen, die auf unterschiedliche Nischen spezialisiert sind, falls sich die Sequenzen zwischen den Gruppen unterscheiden. Diese Theorie wird durch nachgewiesene Zusammenhänge von Umwelteinflüssen mit relativen Häufigkeiten von Genotypen einer Art unterstützt.
[Wilmes *et al.* (2008)]

Nicht-klonale Modelle: Beide Modelle sind klonal, weil sie Rekombinationen vernachlässigen. In mikrobiellen Proben werden oft hohe Rekombinationsraten beobachtet. Durch Rekombinationen können mehrere Sequenzabschnitte desselben Chromosoms getrennt von einander selektiert werden. Das führt zu einer höheren Heterogenität als unter den klonalen Modellen zu erwarten wäre. Auffällige Cluster sind viel seltener zu beobachten, wenn Rekombinationsraten größer als ein Viertel der Mutationsrate werden. Wenn sehr viele Rekombinationsereignisse in einer Population vorliegen, ist es problematisch phylogenetische Abstammungslinien zu identifizieren. Umgekehrt werden Probleme bei der Erstellung phylogenetischer Bäume als Hinweise auf Rekombinationsereignisse verstanden. Rekombinationsraten nehmen mit zunehmender Inhomogenität der beteiligten Sequenzen ab.
[Wilmes *et al.* (2008)]

B.2.6. Mitochondrien

Mitochondrien sind die membran-umgebenen Bestandteile eukaryotischer Zellen, die für die Sauerstoffverwertung verantwortlich sind. Sie enthalten ihre eigene DNA und sind mehrfach in den Zellen enthalten.
[Bullerwell & Lang (2005)]

Eine gewöhnliche tierische Zelle enthält hunderte bis über tausend Mitochondrien und jedes Mitochondrion 5-10 mtDNA Kopien. Mitochondrien fusionieren regelmäßig miteinander und durchmischen so ihre DNA. Ihre DNA (mtDNA) ist innerhalb eines Organismus weitgehend identisch.
[Storch *et al.* (2013)]

Sie kodiert nur für 5 bis ca. 100 Proteine. Die meisten ihrer Proteine werden stattdessen im Zellkern kodiert und als fertige Proteine in die Mitochondrien importiert. Die Mutationsrate von mtDNA ist höher als die der nuklearen DNA ((d.h. DNA des Zellkerns)). Da Mechanismen existieren, die mtDNA in den Kern einschleusen können und Mitochondrien ihrerseits viele Biomoleküle importieren können, ist die Reduktion der mtDNA evolutionär begünstigt. Außerdem werden viele Aufgaben bereits durch die nukleare DNA abgedeckt und die entsprechende mtDNA kann ohne Schaden verloren gehen. Ein extremes Beispiel ist die tRNA für das AUA-Codon. Einige Pilze sind dazu übergegangen diese von der Zelle zu importieren, obwohl sie diese auch selbst kodieren. Einige kodieren diese tRNA nicht mehr und sind

auf den Import angewiesen. Einige brauchen diese tRNA gar nicht, da sie keine Gene mehr haben, deren Sequenz dieses Codon enthält.
[Bullerwell & Lang (2005)]

Es wurden auch Genvereinfachungen beobachtet, die z.B. zum Wegfall großer räumlicher Bereiche bei tRNAs geführt hat. Von wenigen konservierten Bereichen abgesehen, ist mtDNA sehr divers. Zwischen mtDNA-Länge und der Anzahl der Gene scheint es keinen Zusammenhang zu geben: Die meisten Längenunterschiede stammen aus repetierenden Sequenzen in Introns oder zwischen Genen. Selbst innerhalb einer Gattung kann die mtDNA-Länge deutlich variieren. Plasmide sind in Pilzen häufig zu finden, treten aber auch in einigen Pflanzen und Protisten auf. In Pilzen sind sie meist linear. Wenn ein lineares Plasmid in mtDNA eingebaut wird, wird die mtDNA auch linear. Die meisten Mitochondrien enthalten mehrere lineare Chromosome, die sich zu einer kreisförmigen Struktur zusammenlagern, so dass lange Zeit angenommen wurde, es würde sich um ein einziges geschlossenes Chromosom handeln. Es gibt auch Mitochondrien, in denen die mtDNA in mehrere kreisförmige Strukturen organisiert ist. Das mitochondriale Genom kann über mehre Mitochondrien einer Zelle verteilt vorliegen. Die Anzahl der Kopien eines Gens ist oftmals abhängig von Wachstumsbedingungen und liegt zwischen zehn und mehreren tausend. Die hohe Anzahl an Kopien ermöglicht Mutationen ohne sofortige negative Auswirkung, was zu der erhöhten Rate akzeptierter Mutationen beiträgt.
[Burger *et al.* (2003)]

B.2.7. Verteilung von SNPs

SNPs sind nicht gleichmäßig im Genom verteilt. Hypervariable Regionen werden häufig Gen-Inseln genannt und enthalten deutlich höhere Anteile an neuen Genen. In Bakterien weist der "clusterd regularly interspaced short palindromic repeats" (CRISPR) Lokus die meiste fein-skalierte Heterogenität auf. Er ist für einen Teil der Resistenz gegen Bakteriophagen verantwortlich.
[Wilmes *et al.* (2008)]

B.2.8. Mutationen in Viren

Mutationen lassen sich in Viren relativ leicht untersuchen.

Mutation zu RNA-Basen: In DNA mutiert die Base Cytidin gelegentlich zu Uracil. Es existieren Mechanismen, die entsprechende Stellen finden und korrigieren. Wenn sie nicht korrigiert werden, wird bei der nächsten Replikation an diesen Stellen "A" anstatt "G" im komplimentären Strang eingebaut. Das HIV-1 Virus verpackt gewöhnlich ein Korrekturenzym des Wirts gegen diese Mutation mit in seine Kapsel. Geschieht dies nicht, erhöht sich seine Mutationsrate, je nach Zelltyp und -status, auf das 4 bis 18-fache.
[Sanjuan & Domingo-Calap (2016)]

Mutation und Rekombination in RNA-Viren hängen sehr von der Genauigkeit der Polymerase ab. Die Genauigkeit wird durch gut definierte RNA-Sekundärstruktur-Elemente vermindert. Sekundärstrukturen können auch zum Wechsel der Vorlage führen. Auf diese Weise funktioniert Rekombination bei RNA-Viren.
[Sanjuan & Domingo-Calap (2016)]

B.2.8.1. Mutationsraten in Viren

Mutationsraten unterscheiden sich zwischen Viren. Chromosome einsträngiger Viren sind instabiler gegenüber chemischen induzierten Schäden. Die Anzahl der erzeugten Mutationen bei der Replikation des Chromosoms eines DNA-Virus scheint relativ unabhängig von der Größe bei etwa bei 0.003 zu liegen. Die Mutationrate von RNA-Viren liegt zwischen 10^{-6} und 10^{-4} pro Nukleotid und Generation. Der Unterschied zu DNA-Viren ($10^{-8} - 10^{-6}$ pro Nukleotid und Generation) liegt vor allem daran, dass nur bei sehr wenigen RNA Viren eine Fehlerkorrektur durch Exonukleasen stattfindet. (Exonuklease sind Enzyme, die Nukleotide vom Ende der DNA- oder RNA-Kette abspalten [Knippers (1997)]). Coronaviren, die größten (30000-33000 bp) RNA-Viren, sind die einzigen bekannten RNA-Viren, die Fehlerkorrekturmechanismen entwickelt haben.
[Sanjuan & Domingo-Calap (2016)]

Einfluss auf Mutationsraten: Mutationsraten unterliegen Selektion und werden auf verschiedene Weisen gesteuert durch:

- Genauigkeit, mit der Polymerasen bei der Replikation Nukleotide einbauen
- Sequenz
- Sekundärstruktur
- Bedingungen innerhalb der Zelle
- Replikationsmechanismus
- Fehlerkorrektur während der Replikation (bei DNA-Viren)
- Reparatur nach der Replikation (bei doppelsträngigen DNA-Viren)
- Virus-kodierte Diversitäts-erzeugende Retro-Elemente (DGR)
- Wirts-kodierte Deaminasen (bei RNA-Viren in Tieren)

[Sanjuan & Domingo-Calap (2016)]

Mutationen durch Wirts-kodierte Deaminasen: 98% der Mutationen in HIV-1 werden durch sequenzabhängige Deaminasen der Wirtszelle verursacht, die als Abwehrmaßnahmen tierischer Zellen gegen RNA-Viren entstanden sind. Das Virus HIV-1 ist das am schnellsten mutierende bekannte Virus. Es realisiert gewöhnlich in jedem infizierten Patienten an jedem Tag alle einzelnen Basenaustauschmöglichkeiten. Die meisten der editierten Viren sind nicht funktionsfähig. Einige Viren haben Proteine entwickelt, die den Abbau von Deaminasen anregen.
[Sanjuan & Domingo-Calap (2016)]

B.2.8.2. Ortsspezifische Mutationen

In DNA-Viren und Bakterien existieren Mechanismen, die Mutationsraten an speziellen Regionen erhöhen und so optimale Diversität an den entscheidenden Stellen bei geringerer Gesamt-Mutationsrate erreichen können. RNA-Viren scheinen dazu nicht fähig zu sein. Die Raten für akzeptierte Mutationen zeigen sich hingegen auch bei ihnen erhöht – zum Beispiel in Bereichen, die das Immunsystem eines Wirts zum Erkennen benutzt.
[Sanjuan & Domingo-Calap (2016)]

Methylierung: Dem schnellst mutierenden DNA-Virus (Bakteriophage $\Phi X174$) fehlen methylierbare Sequenzmuster, an denen Korrekturenzyme den Originalstrang erkennen könnten, obwohl davon 20 durch Zufall zu erwarten wären. Er mutiert mit 10^{-6} Mutationen pro Nukleotid und Generation um 3 Größenordnungen schneller als das Bakterium *E. coli*. Wenn die Methylierung nur an bestimmten Stellen fehlt, ist die Mutationsrate dort ortsspezifisch erhöht.
[Sanjuan & Domingo-Calap (2016)]

Diversitäts-Erzeugende Retro-Elemente: In großen DNA-Bakteriophagen kann die Mutationsrate ortsspezifische durch Diversitäts-Erzeugende Retro-Elemente (DGRs) gesteigert werden, die sich üblicherweise in Bereichen finden, die für das Anheften an Bakterienzellen verantwortlich sind. DGRs enthalten einen Bereich der eine konservierte Sequenz wiederholter Motive, eine variable Region und ein Gen, das eine reverse Transkriptase kodiert, enthält. (Reversive Transkriptasen sind Enzyme, die RNA in DNA übersetzen.) Diese reverse Transkriptase ließt das Transkript der konservierten Sequenz ab und erzeugt eine fehlerhafte DNA Sequenz daraus. Diese wird auf momentan unbekannte Weise in die variable Region integriert. DGRs existieren auch in Plasmiden, Chromosomen von Prokaryoten und Viren von Archaeen. DGRs selbst wurden nicht in eukaryotischen Viren beobachtet – aber ähnliche Sequenzen.
[Sanjuan & Domingo-Calap (2016)]

Kopienzahlvariationen: Eukaryotische DNA-Viren besitzen andere Methoden um die Mutationsrate ortsspezifisch zu erhöhen, die über Rekombinationen funktionieren. Das Vaccinia Virus enthält beispielsweise CNVs.
[Sanjuan & Domingo-Calap (2016)]

B.3. Ökologie

Next Generation Sequencing von 16S rRNA Genen revolutionieren den Fachbereich mikrobielle Ökologie [Wilmes *et al.* (2008)].

Der Wissenschaftszweig Ökologie beschäftigt sich mit den Prinzipien, die Verbreitung und Abundanz von Organismen bestimmen. Natürlichen Selektion ist daher auch eine ökologische Theorie. Oft sind aktuelle Lebensgemeinschaften nicht optimal aufeinander abgestimmt und unter anderen Bedingungen entstanden.
[Begon *et al.* (2016)]

B.3.1. Lebensgemeinschaften:

Die "Lebensgemeinschaft" wird durch alle Lebewesen an einem gemeinsamen Ort gebildet. Der Ort ist das "Habitat" der dort lebenden Organismen. Der Begriff "Nische" (s.u.) wird in der Umgangssprache häufig mit "Habitat" verwechselt. Gewöhnlich bilden mehrere unterschiedliche Habitate zusammen einen "Lebensraum". Die Lebensgemeinschaft ergibt zusammen mit den abiotischen Umweltfaktoren das "Biotop". Lebewesen derselben Art an einem Habitat bilden eine "Population". Veränderungen der Populationsstruktur (insbes. die Anzahl der Individuen) werden als "Populationsdynamik" bezeichnet [Sauermost & Freudig (1999)]. Die Populationsdynamik vieler parasitärer Bakterien und Viren verläuft zyklisch.
[Begon *et al.* (2016), Nentwig *et al.* (2011)]

Planstelle: Populationen besetzen "Planstellen" in einer Lebensgemeinschaft. Eine Planstelle braucht nicht besetzt zu sein. Sie kann in der Regel von verschiedenen Arten besetzt werden, die nicht nahe-verwandt zu sein brauchen. Dies nennt man "Stellenäquivalenz". An einem Ort zu einer Zeit wird sie in der Regel jedoch nur von einer Art besetzt. Gemeinschaften sind "ungesättigt", wenn sie offenen Planstellen besitzen und "gesättigt" falls alle belegt sind. Offene Planstellen beschleunigen Evolution – besonders nach Massenaussterben. Mit zunehmendem Artenreichtum wird normalerweise jede weitere Art für die Funktionsfähigkeit der Gemeinschaft weniger wichtig, aber eine hinzukommende Art kann auch neue Planstellen schaffen.
[Nentwig *et al.* (2011)]

Der Artenreichtum einer Gemeinschaft steigt mit größerer Ressourcenvielfalt, stärkerer Spezialisierung der Arten, höher Überlappung der Ressourcennutzung verschiedener Arten und weniger freien Lücken. Die Größe einer Population kann durch die Verfügbarkeit von Ressourcen und das Vorhandensein von Prädatoren oder parasitären Arten limitiert sein. Die relativen Artenhäufigkeiten sind sehr unausgeglichen in den meisten Gemeinschaften: nur wenige Arten kommen häufig vor und die meisten selten. Eine häufig vorkommende Art wird "dominant" genannt. Ist eine Art konkurrenzfähig und kommt gut mit den Umweltbedingungen zurecht, wird sie häufig vorkommen. Prädation kann den Artenreichtum erniedrigen, wenn Beute ausstirbt aber auch erhöhen, falls die Beute dominant ist und soweit reduziert wird, dass Ressourcen für weniger konkurrenzfähige Arten verfügbar werden. Parasiten können die Populationsstruktur ihrer Wirte genauso beeinflussen wie Räuber die ihrer Beute [Lucius *et al.* (2018)]. Der Artenreichtum ist bei mittlerem Prädationsdruck am höchsten und kann sich durch Immigration vergrößern. Der Standort bestimmt den regional zur Verfügung stehenden Artenpool.
[Begon *et al.* (2016), Nentwig *et al.* (2011)]

B.3.1.1. Ökologische Nische

Ökologische Nische ist eine Abstraktion darüber, welche Funktion eine Art in der Gemeinschaft ausführen kann und welche Ansprüche sie an ihre Umwelt stellt. Arten, die auf wenige Ressourcen spezialisiert sind, haben eine geringe "Nischenbreite". Zwei Arten, die viele gemeinsame Ressourcen verwenden, haben eine hohe "Nischenüberlappung". Arten, die ähnliche Nischen (bzgl. Ressourcen) haben werden zu einer "Gilde" zusammengefasst. Nah-verwandte Arten haben zwangsläufig ähnliche Nischen. In einer konkreten Gemeinschaft ergibt sich für jede Art ihre "realisierte Nische". Diese kann von der theoretischen individuellen "fundamentalen Nische" abweichen.
[Begon *et al.* (2016), Nentwig *et al.* (2011)]

Interspezifische Konkurrenz: Wenn verschiedene Arten dieselben knappen Ressourcen verwenden, stehen sie in "interspezifischer Konkurrenz". Beeinflussen sich die Populationen dadurch, dass sie die Ressourcen verbrauchen, wird von "Ausbeutungskonkurrenz" gesprochen. Wenn sie um Zugang zu Ressourcen kämpfen, wird die Konkurrenz "Interferenz" genannt. Interspezifische Konkurrenz wirkt auf Gilden und beeinflusst so meist Populationsdynamik, Verbreitung der Arten und Zusammensetzung der Gemeinschaften. Die Habitate eines Lebensraums bestehen normalerweise aus günstigen und ungünstigen Habitaten, von denen einzelne oft nur vorübergehend an unvorhersehbaren Orten und Zeiten existieren. Dies beeinflusst die Wirkung der Konkurrenz. Das "Konkurrenzausschlussprinzip" besagt, dass, wenn zwei Arten einer stabilen Lebensgemeinschaft versuchen dieselbe Nischen zu belegen, die konkurrenzschwächere verdrängt wird. Sie kann daraufhin andere Ressourcen verwenden oder an einem (evtl. nur geringfügig) anderen Ort oder einer anderen Zeit aktiv sein. Selbst bei nur ähnlichen Nischen kann ein Konkurrenzausschluss entstehen. Konkurrenzversuche werden häufig zwischen Stämmen einer Bakterienart durchgeführt. Der Anteil

aktuell noch bestehender interspezifischer Konkurrenz in Lebensgemeischaften wird vermutlich aufgrund auffälliger Beispiele überschätzt. Evolution reduziert langfristig interspezifische Konkurrenz und verringert damit Nischenüberlappungen in Lebensgemeinschaften. Dies wird "Merkmalsverschiebung" genannt. Insbesondere hat natürliche Selektion bisher Individuen begünstigt, deren Eigenschaften sich möglichst nicht nachtteilig auf Individuen der eigenen Population auswirken. Nah-verwandte Bakterien neigen dazu in denselben Proben gemessen zu werden [Lozupone *et al.* (2012)]. Da Arten derselben Gattung phylogenetisch eng miteinander verwandt sind, deuten Lebensgemeinschaften, in denen nicht wenige Gattungen mit mehreren Arten vertreten sind, darauf hin, dass dort Umweltbedingungen eine größere Rolle spielen als Konkurrenz. Bei "ausbeutervermittelter Coexistenz" wird beispielsweise die Dominanz der konkurrenzstärksten Art dadurch reduziert, dass sie als Nahrungsquelle genutzt wird. Das gilt auch wenn die konkurrierenden Arten ebenfalls als Nahrungsquelle verwendet werden. Ausbeuter können auch parasitäre Lebewesen sein.
[Begon *et al.* (2016), Nentwig *et al.* (2011)]

Es wird angenommen, dass nah-verwandte Arten unter ökologischen Gesichtspunkten ähnlich sind und stärker miteinander konkurrieren. Weniger verwandte Arten sollten häufiger miteinander kooperieren. Allerdings zeigen sich einige Arten, die unabhängig von ihrer Verwandtschaft, generell viele Kooperationen eingehen.
[Venail *et al.* (2014)]

B.3.2. Änderungen in Lebensgemeinschaften

B.3.2.1. Stabilität von Umwelten und Gemeinschaften

Selbst in einer stabilen Gemeinschaft werden Arten ausgetauscht. Beim "Species-Turnover" wandern neue Arten in die Lücken ein, die aussterbende hinterlassen. Eine Umwelt gilt als stabil, wenn sie nur wenig unvorhersehbare Schwankung aufweist. Zyklische Veränderungen sind daher als stabil anzusehen.
[Nentwig *et al.* (2011)]

Mit zunehmendem Artenreichtum, wird eine Gemeinschaft stabiler, wobei jede einzelne Population sogar möglicherweise leicht an Stabilität verliert. Gemeinschaften, die nach Störungen schnell wieder in ihren ursprünglichen Zustand zurückkehren, werden "resilient" genannt. Eine Gemeinschaft, die sich durch Störungen kaum verändern lässt, ist "resistent".
[Begon *et al.* (2016)]

Positives Feedback ist oft verantwortlich für Änderungen in Gemeinschaften, kann aber auch Gruppen in einer Gemeinschaft stabilisieren, die sich gegenseitig fördern. Negatives Feedback erhöht die Diversität und somit meist auch die Resilienz einer Gemeinschaft.
[Lozupone *et al.* (2012)]

Menschlicher Krankheitsverlauf scheint oft besser durch die Instabilität mikrobieller Gemeinschaft vorhersagbar zu sein als durch einzelnen Pathogene [Knight *et al.* (2018)].

Zyklische Veränderungen von Populationsdichten in Wirt-Parasit-Beziehungen: Eine Ursache für zyklische Veränderungen von Populationsdichten, liegt in zeitlich-verzögerten Auswirkungen von Wechselbeziehungen zwischen Arten. So erleichtern hohe Wirtsdichten die Ausbreitung der Erreger von Infektionskrankheiten. Diese erhöhen die wiederum die Mortalitätsrate der Wirte und reduzieren dessen Abundanz.
[Nentwig *et al.* (2011)]

Für Parasiten rentieren sich insbesondere Anpassungen an den häufigsten Wirtsgenotyp. Somit entstehen Vorteile für Mitglieder der Wirtsart mit selteneren Genkombinationen. ((Siehe: Disruptive Selektion)). Wenn diese häufiger werden, können sich auch Parasiten, die auf den entsprechenden Typ spezialisiert sind, besser ausbreiten. Auf diese Weise können Schwankungen innerhalb der Genotypenhäufigkeiten einer Art auftreten und ein großes Pan-Genom erhalten bleiben.
[Lucius *et al.* (2018)]

r/K-Selektion: Populationen wachsen unter optimalen Bedingungen zunächst exponentiell an. Die Wachstumskurve flacht ab, da sich die Konkurrenz zwischen den Individuen ("intraspezifische Konkurrenz") der Art um Ressourcen erhöht. In einer gegebenen Umwelt ist die "Kapazität" einer Population ihre theoretisch maximale Anzahl an Individuen. In einer stabilen Umwelt kann diese zumindest annähernd erreicht werden. Die Mitglieder einer Art konkurrieren auch mit anderen Arten um Ressourcen. Ihre wichtigste Eigenschaft ist dann Konkurrenzfähigkeit. In der Situation spricht man von "K-Selektion". Konkurrenzfähige Arten werden "K-Strategen" genannt. Habitate, die sie begünstigen sind "K-selektierend". In einer instabilen Umwelt ist hingegen die Anzahl der Nachkommen ausschlaggebend. Man spricht von "r-Selektion", "r-Strategen" und "r-selektierenden Habitaten. Da sich diese Arten in neuen Situationen schnell ausbreiten können, werden sie auch "opportunistisch" genannt. Gemeinschaften mit einem hohen Anteil an r-Strategen sind besonders resilient. Individuelle r-Strategen sind stärker von bereits kleinen Umweltschwankungen beeinträchtigt, ihre Populationen sind aber weniger einheitlich. Die r und K Einteilung ist ein sehr hilfreiches, einfaches und weitreichend gültiges Konzept. Einige Arten können sowohl K- als auch r-Strategie anwenden.
[Nentwig *et al.* (2011), Begon *et al.* (2016)]

Organismen mit kleinem Genom können ihr Genom einfacherer reproduzieren und sollten einen schnelleren Reproduktionszyklus aufweisen [Lucius *et al.* (2018)].

Ökologische Sukzession bezeichnet die Erstbesiedlung eines Ortes oder seine Wiederbesiedlung nach einer Störung. Alte Arten können dabei verloren gehen. Der Prozess kann stufenweise ("sukzessiv") ablaufen, wobei einige neue Arten nach einem Übergangszeitraum wieder verschwinden. Die Veränderungen sind nicht-saisonal und streben einen Endzustand im Gleichgewicht an. Zu Beginn einer Sukzession sind Artenreichtum und Biomasse gering, werden aber im Laufe der Sukzession größer. Der Artenreichtum kann zu einem mittleren

Zeitpunkt am größten sein, wenn die Arten aus konsekutiven Sukzessionsstadien gleichzeitig auftreten. ((Größte Diversität entspricht also nicht zwangsläufig dem best-angepasstem Zustand.)) Während die Zusammensetzung der Arten anfänglich sehr zufällig ist und von r-Strategen dominiert wird, setzten sich später meist gut angepasste K-Strategen durch. Die frühen Arten können anderen Arten helfen oder deren Ansiedlung verhindern, indem sie die Umweltbedingungen anpassen. Außerdem können sie die Stelle besser gegen spätere Arten verteidigen.

Häufig treten in Gemeinschaften neue Störungen auf, bevor sich ein stabiler Zustand bilden kann. Der Artenreichtum sollte bei einer mittleren Störungshäufigkeit am größten sein. Oft befinden sich verschiedene Habitatfragmente einer Gemeinschaft in unterschiedlichen Sukzessionsstadien.
[Nentwig *et al.* (2011), Begon *et al.* (2016)]

B.3.2.2. Wanderung

Individuen können ihren Lebensraum wechseln. Aufgrund von Immigration können sich Individuen auch an Orten aufhalten, an denen für sie kein Populationswachstum möglich ist. Sie emigrieren aus Bereichen mit, für sie besseren, Bedingungen und größerer Population. Emigrationshäufigkeiten hängen von Art und Umwelt ab. Immigration wird in der Regel durch "biotischen Widerstand" der etablierten Populationen erschwert. Ihm liegt "diffuse Konkurrenz" zugrunde – höhere Artenanzahl führt tendenziell zu weniger freien und kleineren Nischen. Eine immigrierende Art kann sich nur etablieren, wenn für sie eine freie passende Planstelle existiert oder sie die realisierten Nischen der etablierten Arten verändern kann. Manchmal können sich einwandernde Arten nach vielen Generationen durch Anpassungen etablieren.
[Nentwig *et al.* (2011)]

Immigration kann auch vor Feinden schützen und eine Art sogar dominant werden lassen, wenn die Bedingungen in dem neuen Habitat für die Antagonisten deutlich schlechter ist [Lucius *et al.* (2018)].

B.3.3. Ressourcenvielfalt

Wie aus einem Ökotyp Modell (Abschnitt B.2.5) zu erwarten, nutzen artenreiche Gemeinschaften begrenzte Ressourcen besonders effizient, da unterschiedliche Arten auf ihre Nische spezialisiert sind. Störungen von Gemeinschaften erleichtern meist die Ansiedlung fremder Arten. Ein ausgiebiges Nahrungsangebot reduziert häufig die Diversität eines Ökosystems, da wenige schnell wachsende Arten in interspezifischer Konkurrenz zu anderen stehen.
[Lozupone *et al.* (2012)]

B.3.4. Systemkomplexität und taxonomische Aussagekraft

Selektion basiert auf Eigenschaften und Umwelteinflüssen und verändert die Zusammensetzung von Gemeinschaften. Der Zusammenhang zwischen Eigenschaft und Verwandtschaft reduziert sich durch lateralen Gentransfer. Eigenschaften komplexerer Systeme, die auf dem Zusammenspiel vieler Proteine basieren, neigen dazu lateralen Gentransfer zu vermeiden. Änderungen von Umwelteinflüssen die sehr konservierte Eigenschaften betreffen, sollten die Zusammensetzung auf den oberen gröberen taxonomischen Ebenen beeinflussen. Die Ebene auf der Abundanzunterschiede festgestellt werden, kann Informationen über die Konserviertheit der betroffenen Eigenschaften liefern.

Das Mikrobiom von Patienten mit chronisch-entzündlichen Darmerkrankungen unterscheidet sich von gesunden Patienten in den hohen taxonomischen Ebenen Klasse und Stamm. Die Pathogenität eines Bakteriums unterliegt hingegen häufig lateralem Gentransfer und scheint wenig konserviert zu sein. Gegen welche Phagen ein Bakterium resistent ist, hängt weitgehend von SNPs ab. Das steht einer Konservierung auch diese Eigenschaft entgegen.

Eigenschaftsunterschiede, die erst auf Ebenen konserviert sind, die unterhalb der Ebene liegen, auf der untersucht wird, können in Studien übersehen werden. Im menschlichen Darm scheinen mikrobielle Eigenschaften in Bezug auf Nahrungsquellen ungefähr auf Gattungsebene konserviert zu sein.
[Martiny *et al.* (2015)]

Bemerkung 52:
Wenn Anhand der taxonomischen Ebene Rückschlüsse auf die Konserviertheit einer Eigenschaft gezogen werden sollen, muss darauf geachtet werden, dass diese Einschätzung nicht bloß Anhand weniger Arten getroffen wird. Der PC-Test reagiert beispielsweise sehr sensitiv auf starke Effekte in einzelnen Dimensionen.

B.3.5. Phagen und Bakterien

Die weltweite Anzahl an Phangen wurde auf 10^{31} geschätzt und die durchschnittliche Anzahl der von ihnen verursachten Infektionen in Bakterien auf 10^{23}. Eine Bodenprobe von 10 Gram enthält ungefähr 10^{10} Bakterien und 10^{7} verschieden Bakterien-Arten.
[Medini *et al.* (2005)]

Phagen und Resistenzmechanismen regulieren bakterielle Populationen in den meisten Ökosystemen. Wahrscheinlich können Phagen ihrerseits Kontermechanismen gegen bakterielle Phagenresistenz entwickeln, ohne dass ihre Fitness bezüglich Bakterien ohne entsprechende Resistenzmechanismen deutlich absinkt. In den meisten Ökosystemen unterliegen Phagen und Bakterien einer zyklusförmigen Co-Evolution.
[Labrie *et al.* (2010)]

Phagen-moderierte Human-Pathogenität von Bakterien: Einige Bakterien verändern oder erlangen, wenn sie von entsprechenden Phagen infiziert werden, die Fähigkeit menschliche Zellen zu schädigen. Ihre medizinische Bedeutung hängt von ihrem eigenen Infektionsstatus ab.
[Wagner & Waldor (2002)]

B.3.6. Heterogenität unter Artenebene

In mikrobiellen Gemeinschaften sind feinskalierte genetische Unterschiede innerhalb Populationen derselben Art üblich, die sich manchmal auf Funktionen auswirken. Die Inner-Populations-Variation unterscheidet sich deutlich zwischen Populationen im selben Lebensraum. Bei einigen Populationen sind nur 0.08% SNPs zu beobachten, während es in anderen bis zu 2.2% SNPs sind. Marker-Gen Analysen reichen für solche Betrachtungen nicht aus, da Variationen auch außerhalb des Gens liegen.
[Wilmes *et al.* (2008)]

B.3.7. Pan-Genom

Viele Bakterien derselben Art tragen nicht nur Gene, die sich an einigen Stellen unterscheiden, sondern auch welche, die ihre Artgenossen gar nicht haben. Zur Beschreibung einer Art wird deshalb gelegentlich das Pan-Genom verwendet. Dies ist die Gesamtheit der Gene einer Art und setzt sich aus einem Kerngenom, das alle Bakterien dieser Art aufweisen, und weiteren Genen ("Dispensable Genome") zusammen. Die Gene des "Dispensable Genome" können auf bestimmte Nischen ausgerichtet sein. Solche Gene sind gewöhnlich in große genomische Inseln zusammengefasst, vor und hinter denen sich kurze, wiederholte DNA Sequenzen befinden. Diese Gene scheinen häufiger durch Phagen oder Transposonen entstanden zu sein als das Kerngenom und können frei zwischen verschiedenen Stämmen ausgetauscht werden. Sie weisen außerdem einen ungewöhnlich hohen G–C Anteil auf. Gene mit pathogener Bedeutung liegen häufig ebenfalls im Dispensable Genome.

Die fehlende Übereinstimmung zwischen Serogruppe und genetischer Diversität liegt wahrscheinlich teilweise daran, dass Gene für Bakterienoberflächen zum Dispensable Genome gehören. Sequenztypen werden hingegen im Kerngenom bestimmt.

Das Pan-Genom kann um Größenordnungen größer sein als das Genom eines einzelnen Bakteriums. Von dem Bakterium *Vibrio splendidus* sind über 1000 verschiedene Genome bekannt. Die Gene des Pan-Genoms werden ständig innerhalb einer Art und zwischen verschiedenen Arten ausgetauscht.
[Medini *et al.* (2005)]

B.3.8. Kultivierte Mikroben und Microbial Dark Matter

Weniger als 1% aller Mikroorganismen lässt sich kultivieren [Chen *et al.* (2020)]. Die Gesamtheit der noch nicht kultivierten Mikrobenarten wird "microbial dark matter" genannt. Das National Center for Biotechnology Information (`www.ncbi.nlm.nih.gov`) enthält eine fast vollständige Datenbank bisher komplett sequenzierter 16S rRNA Gene. Da diese Sequenzdaten hauptsächlich von amplifizierten Sequenzen stammen, sind viele Verfälschungen zu erwarten. Außerdem sind 16S Sequenzen ungeeignet zur vollständig korrekten taxonomischen Zuordnung. Werden diese Umstände ignoriert, erkennt man mit dieser Datenbank:

- Die meisten Mikrobenarten, die im Menschen vorkommen, wurden bereits in Laboren kultiviert.
- Der Mensch ist das einzige Ökosystem für das dies zutrifft.
- Ungefähr ein Viertel prokaryotischer OTUs ist auf Familienebene mit keinem kultivierten Organismus verwandt.
- 1657 vollständig sequenzierte Bakterien-Genome weisen im Durchschnitt 3.8 Kopien des 16S rRNA Gens auf.
- Bei 79 Archaeen sind es durchschnittlich 1.8 Kopien.
- Zu unkultivierten Prokaryoten lagen keine vollständig sequenzierten Genome vor.

[Lloyd *et al.* (2018)]

Die Kultivierbarkeit humaner-Mikroben wird gelegentlich überschätzt, so galten gesunde Lungen lange Zeit als steril, was durch kultur-unabhängige Verfahren inzwischen widerlegt wurde [Tiew *et al.* (2020)]. Obwohl auch Darmbakterien mit heutigen Mitteln zum großen Teil kultiviert werden können, zeigen sich in Mikrobiomanalysen auch einige versuchsrelevante Bakterien, für die es noch nicht möglich ist [Lozupone *et al.* (2012)]. Im Jahr 2002 waren über die Hälfte menschlicher Darmbakterien noch nicht kultivierbar [Shanahan (2002)].

B.3.9. Der Darm als Ökosystem

In einem gewöhnlichen menschlichen Darm leben über 1000 verschiedene mikrobielle Arten. Neben Bakterien kommen auch Archaeen, Eukaryoten (hauptsächlich Hefe) und Viren (hauptsächlich Phagen) vor. Untersuchungen von Bakterienkulturen schienen zu zeigen, dass die meisten Arten in allen gesunden Erwachsenen vorkommen – dies wurde jedoch durch sequenzbasierte Verfahren widerlegt. Die Mikrobengemeinschaft variiert zwischen Zeitpunkten und noch mehr zwischen Menschen. Zeitliche Schwankungen liegen insbesondere bei Erkrankungen und bei kleinen Kindern vor. Insbesondere während der ersten 3 Lebensjahre findet eine ökologische Sukzession im Darm statt.
[Lozupone *et al.* (2012)]

Die Mikroben leben in mutalistischer Beziehung zu ihrem Wirt. Sie erhalten konstante Temperatur und pH-Wert und Ressourcen für ihr Wachstum unter anaeroben Bedingungen.

Die Wirte erhalten die Möglichkeit mehr Nahrung zu verwerten zu können und gesundheitliche Vorteile.
[Begon *et al.* (2016)]

Gesundheit: Die Gemeinschaft schützt vor Pathogenen im Darm, übernimmt Verdauungsaufgaben und beeinflusst bei Kindern die Entwicklung des Immunsystems. Störungen des Gleichgewichts der Gemeinschaft werden mit Fettleibigkeit, Unterernährung, Darmerkrankungen, neurologischen Störungen und Krebs in Verbindung gebracht. Mikrobiome von Menschen mit identischer Krankheit unterscheiden sich häufig stärker voneinander als die von Gesunden – wobei diese Erkenntnis hauptsächlich auf Studien aufbaut, die ehemaligen Konsum von Antibiotika nicht berücksichtigt haben. Einige Charakteristiken gestörter mikrobieller Gemeinschaften in Erwachsenen ähneln denen ungestörter in Kindern.

Der Zusammenhang zwischen Gesundheitszustand und der Mikrobengemeinschaft erfordert einen ökologischen Blickwinkel und lässt sich nicht durch betrachten der Pathogene verstehen. Mikrobielle Darmgemeinschaften sind normalerweise resistent gegen Besiedlungsversuche fremder Mikroben und andere Einwirkungen. Medikamente und Diäten (einschließlich Prä- und Probiotika) können aber auch dauerhafte Veränderungen bewirken. Eine 10-Tage Diät verursacht Änderungen am Mikrobiom, die kleiner sind als zwischenmenschliche Unterschiede. Langzeitdiäten oder Antibiotika können zu neuen stabilen Gemeinschaften führen. Verschiedene Antibiotika wirken unterschiedlich auf die mikrobiellen Gemeinschaften verschiedener Menschen. Bei Fettleibigkeit ist sowohl die Artenvielfalt als auch die Anzahl funktioneller Gene reduziert. Keimfrei aufgezogene Mäuse werden eher adipös, wenn ihnen die Mikrobiomgemeinschaft fettleibiger als schlanker Mäuse eingesetzt wird.
[Lozupone *et al.* (2012)]

Funktionale Redundanz besteht, weil Mikrobiome auf Gen-Ebene ähnlicher sind als auf phylogenetischer. Dieselben Aufgaben werden von unterschiedlichen Arten übernommen. Häufig erfüllen unterschiedliche Mikroben ähnliche Funktionen ohne eine nahe Verwandtschaftsbeziehung zu haben. Mitunter kann im Darm dieselbe Funktion von einem Bakterium und einer Archaee ausgeführt werden. Antibiotika-Einnahme kann abundante aber darauf sensitive Mikroben reduzieren, während deren Funktionen von anderen Mikroben weitergeführt werden, die sich nun eventuell stärker vermehren können. Konkurrierende Mikroben produzieren manchmal gegeneinander antimikrobielle Substanzen. In Mausstudien zeigte sich ein positiver Zusammenhang zwischen der Häufigkeit eng-verwandter Arten und der Anfälligkeit für Invasionen fremder Bakterien.
[Lozupone *et al.* (2012)]

B.3.10. Die Pflanze als Ökosystem

B.3.10.1. Rhizosphäre

Der Bodenbereich, der Wurzeln von Pflanzen umgibt und von diesen beeinflusst wird, wird "Rhizosphäre" genannt. Die Zusammensetzung von Mikrobengemeinschaften in der Rhizosphäre wird hauptsächlich durch Pflanzenarten und Bodeneigenschaft bestimmt. Es wird angenommen, dass eine diversere Pflanzengemeinschaft auch zu einer diverseren Mikrobengemeinschaft führt. Der Bodentyp hat starken Einfluss auf die mikrobielle Zusammensetzung aber die Pflanze kann noch wichtiger sein. Der Standort bestimmt hingegen die Organismen, die der Pflanze generell zur Verfügung stehen.
[Philippot *et al.* (2013)]

Ernährung und Pathogenabwehr: Mikroben der Wurzelregion sind essentiell für den Stickstoffkreislauf zur Ernährung der Pflanze. Stickstoff ist in natürlichen Böden meist der limitierende Faktor für das Pflanzenwachstum. Mikroben aus der Rhizosphäre werden von abgesonderten Stoffen der Wurzel angezogen und ernährt. Die Zusammensetzung der Stoffe hängt unter anderem von der Art der Pflanze und ihrem Alter ab. Zu den Absonderungen gehören auch antimikrobielle Wirkstoffe, die Pathogene abhalten.

Das Mikrobiome in der Rhizosphäre ist wichtig für Wachstum, Ernährung und Gesundheit der Pflanze. Schützende Effekte durch Mikroben der Rhizosphäre sind am bekanntesten von Böden in denen Pflanzen immun oder gestärkt gegen bestimmte Krankheiten sind. In den meisten dieser Böden, stammt diese Wirkung von Teilen der Mikrobengemeinschaft des Bodens und der Rhizosphäre. Sowohl Pilze als auch Bakterien können systemische Reaktionen der Pflanze auslösen, die üblicherweise gegen mehrere Pathogene und Insekten schützen. Diese natürlichen Krankheits-Kontrollen sind eventuell in Züchtungen verlorengegangen - aufgrund von Düngung, Bodenbearbeitung und Selektion nach Ertrag. Es wird angenommen, dass moderne Züchtung gegen Pflanzeneigenschaften selektiert haben könnte, die wichtig für Mikroorganismen gewesen sein könnten. Mikroorganismen die sich in oder auf Samen befinden, beeinflussen möglicherweise ebenfalls die Zusammensetzung des Mikrobioms der Rhizosphäre der späteren Pflanze. Mikroben innerhalb von Samen können wahrscheinlich auch auf weitere Generationen vererbt werden.
[Philippot *et al.* (2013)]

Bemerkung 53:
Die Pilzdaten, die in dieser Arbeit verwendet werden, stammen aus einem Versuch, in dem ein Bakterium auf die Pflanzensamen aufgebracht wurde, das die Pflanze vor einem Pathogenen Pilz schützen soll, ohne die restliche Pilzgemeinschaft zu negativ zu beeinflussen [Antweiler et al. *(2017)].*

Räumlich- und zeitliche Einflüsse: Die Diversität in der Rhizosphäre ist oft geringer als in angrenzenden Bodenbereichen. Die Rhizosphärengemeinschaft unterliegt räumlichen

und zeitlichen Variationen entlang der Wurzeln, die von Alter und Zustand der Pflanze abhängen. Die Mikrobengemeinschaften entlang der Wurzeln unterscheiden sich von denen an den Wurzelspitzen in ihrem Anteil schnell-wachsender Bakterien. Dies könnte für unterschiedliche Selektionsbedingungen sprechen, wobei die Wurzelspitze eine instabile aber nahrungsreiche Umgebung darstellt, während die anderen Teile stabiler aber mit weniger Nährstoffen versehen sind.
[Philippot *et al.* (2013)]

B.3.10.2. Wood Wide Web

Wurzeln sind der bevorzugte Lebensraum von Bodenpilzen. Das sogenannte "Wood Wide Web" wird von miteinander verbundenen Bodenpilzen und Wurzeln gebildet. Bakterien scheinen ebenfalls an der Schnittstelle beteiligt zu sein. Die Organismen fördern sich gegenseitig, insbesondere durch Nahrungstransfer entlang des Netzwerks. Die häufigste Symbiose von Pflanzen ist die mit Pilzen an ihren Wurzeln. Sie findet bei den meisten Pflanzen statt - sowohl in natürlich Böden als auch in Agrikultur. Durch das vergrößerte Netzwerk können Wasser und essentielle Elemente besser aufgenommen werden.
[Bonfante & Anca (2009), Raven *et al.* (2000)]

B.3.11. Ökologische Kenngrößen und Abstände

Ökologische Kenngrößen: Alphadiversität bezeichnet die Diversität innerhalb einer Gemeinschaft ((also die "Gleichmäßigkeit" und "Vielfältigkeit" einer empirischen Verteilung von Organismen)). Eine häufig verwendete Größe ist der Shannon-Index. Phylogenetische Information kann ebenfalls einfließen, zum Beispiel geschieht dies bei dem Maß "Faith's phylogenetische Diversität".
[Knight *et al.* (2018)]

Definition B.2 (Shannon-Index bzw. Entropie). [Shannon (1948)]
Seien $p_1, \ldots, p_n \in [0, 1]$ mit $\sum_{i=1}^{n} p_i = 1$, dann wird

$$H := -\sum_{i=1}^{n} p_i \log p_i$$

die "Entropie" der Verteilung $\{p_1, \ldots, p_n\}$ genannt. Die Basis des Logarithmus ist nicht genau festgelegt und entspricht einer Messeinheit. In der Ökologie wird normalerweise der natürliche Logarithmus verwendet und vom "Shannon-Index" gesprochen [Begon *et al.* (2016)].

Faith's phylogenetische Diversität bildet die Grundlage für den UniFrac Abstand.

Definition B.3 (Faith's phylogenetische Diversität). [Faith (1992)]
Zu einem gegeben phylogenetischen Referenzbaum ist "Faith's phylogenetische Diversität" definiert als die Summe der Astlängen des Unterbaums, der genau die in der Probe vorkommenden OTUs als Blätter enthält.

Bemerkung 54:
Eine OTU kommt genau dann in einer Probe vor, wenn ihre Häufigkeit dort positiv ist. Dieser Unterbaum enthält die minimale Anzahl nötiger Kanten und innerer Knoten. Die alte Wurzel ist eventuell nicht notwendig. Es können innere Knoten auftreten, die nur ein Kind haben. So ein Knoten könnte auch entfernt werden, wenn die anliegenden Kanten zu einer Kante verschmolzen werden, die so lang ist wie die beiden Kanten zusammen. Wenn aus Mangel an phylogenetischer Information allen Kanten dieselbe Länge zugewiesen wird, wäre darauf zu achten, dass nach dem Verschmelzen nicht mehr alle Kanten dieselbe Länge haben.

Bemerkung 55:
Cadotte et al. *(2010) sind der Auffassung, dass Faith's phylogenetische Diversität Äste zu einem speziellen Knoten mitzählt, der nicht Teil der Probe ist. Es könnte sich um eine weit verbreitete Interpretation dieses Diversitätsindex handeln. Im Zweifelsfall sollte daher darauf geachtet werden, was die jeweils verwendete Software berechnet und als Eingabe erwartet.*

Ökologische Abstände werden in Abschnitt 4.2.1 auf Seite 23 besprochen.

C. Next Generation Sequencing

C.1. Next Generation Sequencing Verfahren

In Mikrobiomanalysen kommen als Messmethode Next Generation Sequencing Verfahren wie zum Beispiel die Pyrosequenzierung [Margulies *et al.* (2005)], Illumina MiSeq [Bennett (2004)], PacBio [McCarthy (2010)] oder Oxford Nanopore [Eisenstein (2012)] zum Einsatz. Mit diesen lassen sich, unter anderem, die Mikrobenverteilung der Wurzelregion einer Pflanze untersuchen und mit der von weiteren Pfanzen vergleichen [Schreiter *et al.* (2014)].

Der Sequenzierungsvorgang wird, je nach Höhe der Bezahlung, einige tausend mal parallel durchgeführt. Verschiedene Messmethoden priorisieren, Sequenzanzahl/Kosten (Illumina MiSeq), Fehlerrate (Illumina MiSeq) und Sequenzlänge (PacBio) unterschiedlich [Metzker (2010)]. Bei aktuell sinkenden Kosten und methodischen Verbesserungen werden die Anzahl der Messungen und die Länge der sequenzierten Sequenzen allgemein größer. Die Sequenzdaten ermitteln sich direkt aus der Konstruktion des neuen DNA-Strangs. Der bioinformatische Arbeitsablauf lässt sich durch modular-aufgebaute Open-Source Programme (Pipelines oder Frameworks) benutzerfreundlich anpassen und durchführen. Hierbei wird momentan insbesondere qiime [Caporaso *et al.* (2010)] benutzt, das Anfang 2018 in seiner zweiten Version erschienen ist.

C.1.1. Sequenziermethoden

Das experimentelle Bestimmen der Sequenz einer DNA-Kette wird “Sequenzierung” genannt. 1997 waren mehr als 95% der in Datenbanken hinterlegten DNA-Sequenzen noch durch die Kettenabbruch-Methode von Frederick Sanger erzeugt worden [Knippers (1997)]. Sequenzieren von DNA war lange Zeit teuer und aufwendig. Das erste weitgehend vollständige menschliche Genom wurde mit einer automatisieren Sanger-Methode sequenziert. Diese Methode gehört noch nicht zum Next Generation Sequencing (NGS). NGS-Verfahren sind schneller und günstiger und eignen sich daher dazu große Datenmengen zu gewinnen. Hierdurch werden Sequenzierungsverfahren für Fachbereiche interessant, die nicht primär an der Basenfolgen eines speziellen DNA-Fragmentes interessiert sind. Bevor eine DNA sequenziert werden kann, muss die entsprechende Probe im Labor vorbereitet werden. Die DNA wird extrahiert und in kleinere Fragmente aufgespalten. Bei Verfahren der ersten und zweiten Generation müssen diese Fragmente vervielfältigt (“amplifiziert”) werden. Dies geschieht in der Regel über PCR. Bei NGS-Verfahren werden die Fragmente mit sogenannten Barcodes

ergänzt – Sequenzen, die eine Zuordnung der DNA zur Probe erlauben. Auf diese Weise können viele Proben gleichzeitig im Sequenzierungsautomat verarbeitet werden, ohne dass die Zuordnung verloren geht. Meist wird auch eine Primersequenz hinzugefügt, damit kommerzielle erhältliche Primer als Startstück für eine DNA-Synthese verwendet werden können. Meist wird ein Adapter an das DNA-Fragment angebracht, mit dem es an Befestigungsstellen auf einer festen Oberfläche gebunden werden kann, die viele solcher Bindungsstellen in geeigneten Abständen enthält. Bei Verfahren der ersten beiden Generationen, sind mehrere Kopien deselben DNA-Fragments notwendig, für die ein gemeinsames (Konsensus-)Signal gemessen wird. Da diese Verfahren darauf beruhen, dass Nukleotide an einen wachsenden DNA-Strang, beginnend an einem Primer, angebaut werden, ist es wichtig, dass diese Anlagerung gut funktioniert. Wenn die Anlagerung nicht bei allen Kopien des Fragments gleich funktioniert, entsteht eine Phasenverschiebung und ein verrauschtes Signal für den weiteren Prozess. Das ist ein Grund, der die Länge einer bestimmbaren Sequenz limitiert. Ein weiter liegt darin, dass die Abspaltung von Farbstoffen, die zum Markieren an Nukleotide gebunden werden können, nicht immer funktioniert. ((Die alten Farbsignale mischen sich dann mit denen der folgenden Nukleotiden.)) DNA-Fragmente lassen sich auch mit zwei Primersequenzen versehen – einem an jedem Ende. Eine Sequenz passt zu einem (Forward-)Primer, der sich an den entsprechenden DNA-Einzelstrang anlagern und dann (von einer DNA-Polymerase) zu dem passenden zweiten Strang des DNA-Fragments verlängert werden kann. Die zweite Sequenz (am Ende des Einzelstrangs) ist identisch zu einem (Reverse-)Primer. Dieser Primer würde zu dem zweiten Strang (des Doppelstrangs) passen und könnte per DNA-Synthese zu einer Kopie des ersten Strangs verlängert werden. Auf diese Weise lassen sich die Sequenzen, je nach Methode, von beiden Seiten sequenzieren. Hierdruch können größere Sequenzlängen mit zuverlässiger Information erreicht werden. Bei Verfahren der dritten Generation, wird für jedes einzelne Moleküle ein einzelnes Signal gemessen. Auf diese Weise kann keine Phasenverschiebung entstehen. Die Fehlerate ist höher, steigt aber nicht so stark mit zunehmender Länge des Fragments. Bei optischen Verfahren kann ein Farbstoff an einem Nukleotid fehlen oder funktionsunfähig sein oder die Sicht zwischen Kamera und Farbstoff kann blockiert sein. Da keine Kopien des Fragments erzeugt werden müssen, entsteht kein Amplifikationsbias.
[Metzker (2010), Shendure & Ji (2008)]

Im folgenden wird eine Auswahl bedeutsamer Sequenziermethoden erklärt. In der Überschrift wird angezeigt zu welcher Generation die jeweils beschriebene Methode zählt.

C.1.1.1. Sangers Kettenabbruch-Methode (First Generation Sequencing)

Aufbau: Sangers Methode basiert auf dem Prinzip der Sequenzierung durch DNA-Synthese. ((Das gilt auch für alle folgenden Methoden. Oxford NanoPore ist ein Beispiel für ein Verfahren ohne DNA-Synthese.)) Einem (im Sinne von vielen identischen) zu sequenzierenden Einzelstrang-DNA-Fragment werden die zur Erzeugung des zweiten Strangs benötigten Enzyme, Primer und die vier Deoxynukleotide dATP, dGTP, dCTP und dTTP hinzugefügt. Dies

wird in vier verschiedenen Lösungen durchgeführt. Jede dieser Lösung enthält zusätzlich genau eine Art von Dideoxynucleotid: ddATP, ddGTP, ddCTP oder ddTTP. Dieses kann genau dann in die entstehende DNA eingebaut werden, wenn das entsprechende Deoxynukleotid auch eingebaut werden kann. Aber, wenn es eingebaut wird, bricht die Synthese ab – daher der Name der Methode.
[Knippers (1997)]

Sequenzierung: Die vier Lösungen werden in einem Elektrophorese-Gel nebeneinander entlang eines elektrischen Feldes laufen gelassen. ((Ein Elektrophorese-Gel ist ein Gel mit Vertiefungen an einem Ende, in die die zu untersuchenden Lösungen geschüttet werden. An dieser Seite und an der gegenüberliegenden werden Elektroden angebracht, die ein elektrisches Feld erzeugen, in dem die Ionen der Lösung beschleunigt werden. Die Ionen laufen dabei, entsprechend ihrer physikalischen Eigenschaften, unterschiedlich schnell durch das Gel.)) Für jede der Lösungen werden so die entstandenen DNA-Ketten ihrer Größe nach getrennt. Je kleiner die Kette ist, desto leichter fällt es ihr, durch das Gel zu laufen und desto weiter entfernt ist sie vom Startpunkt, wenn das elektrische Feld abgestellt wird. ((Die DNA-Ketten werden anschließend sichtbar gemacht.)) Die Lösung mit der kleinsten Kette gibt an, was die erste Base des DNA-Fragmentes ist. Wenn es beispielsweise die Lösung mit ddGTP war, dann ist die erste Base ein Guanin. Iterativ gibt die Lösung mit der nächst-kleinsten Kette die nächste Base der Sequenz an. (Dies darf auch wieder dieselbe Lösung sein.) Damit ist das Funktionsprinzip des Verfahrens von Sanger erklärt.
[Knippers (1997)]

C.1.1.2. Roche/454 Pyrosequencing (Second Generation Sequencing)

Amplifikation: DNA-Fragmente werden als Einzelstränge an extrem kleinen Kugeln (sogenannte Beads) angebracht, die viele räumlich separierte Bindungsstellen aufweisen. Das geschieht unter Bedingungen, die begünstigen, dass nur ein einziges DNA-Fragment an einem Bead anlagert. Die Beads liegen in wässrigen Tropfen, die von einer öligen Lösung umgeben sind. Dies schafft quasi isolierte Reaktionseinheiten, die in derselben Lösung behandelt werden können. In den Tropfen laufen PCR-Reaktionen ab, die dazu führen, dass alle Bindungsstellen mit Duplikaten des jeweiligen DNA-Fragments besetzt werden.
[Metzker (2010)]
Durch die Vervielfältigung der DNA-Fragmente, werden die Signale im Sequenzierungsschritt stark genug, um gemessen werden zu können.

Sequenzierung - Aufbau: Die Beads werden auf einem Träger (PicoTiterPlate) aufgebracht. Dieser besitzt kleine Vertiefungen (Wells) in die jeweils ein Bead eingelassen wird. In diesen Vertiefung laufen die Sequenzierungsreaktionen ab. Unter den Vertiefungen befinden sich Kameras, die Lichtblitzintensitäten aufzeichnen. Über den Wells befindet sich ein Bereich (Durchflusszelle), über den Chemikalien zugesetzt werden. Kleine Beads werden hinzugefügt, an die Enzyme gekoppelt sind, die indirekt Pyrophosphat unter Lichtabgabe

umsetzen. Die Beads bleiben in den Wells, während kleinere Moleküle in sie hinein und wieder hinaus gespült werden können. ((Enzyme werden bei Reaktionen nicht verbraucht und müssten ersetzt werden, wenn sie mit hinausgespült würden.)) DNA-Polymerase bindet an Primer an den immobilisierten DNA-Fragmenten und beginnt DNA zu synthetisieren, sobald ihr das zur aktuellen Position passende dNTP zur Verfügung steht. Beim Anbau eines Nukleotides wird Pyrophosphat (PP_i) von dem Nukleotid abgespalten und freigesetzt. Die Lichtintensität, wird von den Kameras registriert.
[Metzker (2010)]

Sequenzierung: Durch die Durchflusszelle werden in festgelegter Reihenfolge dNTPs gespült. Wenn das gerade hinzugefügte dNTP an die aktuelle Position eines entstehenden DNA Strang eingebaut werden kann, wird Licht entwickelt. Für Sequenzabschnitte von bis zu 6 aufeinander folgenden identischen Basen ist die entstehende Lichtintensität ungefähr proportional zur Anzahl der eingebauten Nukleotide. Da zu jedem Zeitpunkt jeweils nur eine Art Base zugegeben wird, lässt sich aus der Lichtintensität ableiten, ob diese an der aktuellen Position eines DNA-Fragments eingebaut wird und wie lang der gegebenenfalls eingebaute Sequenzabschnitt ist. Aus Wechsel der Basenart folgt iterativ die Sequenz des DNA-Fragments.

[Metzker (2010)]

Sequenzierungsfehler: Insertionen sind die häufigste und Deletionen die zweithäufigste Fehlerart in den so produzierten Sequenzdaten.
[Metzker (2010)]

C.1.1.3. Illumina/Solexa (Second Generation Sequencing)

Illumina ist momentan wegen hoher Sequenzierleistung von bis zu 1.5e12 Basen insgesamt über alle Sequenzen pro Durchlauf und einer geringen Fehlerrate von 0.1-1% die für Shotgun Metagenome meistgenutzte Plattform. ((Shotgun Sequenzierungen vervielfältigen DNA, zerlegen die Kopien in zufällige Fragmente, sequenzieren diese und setzen die Sequenzinformation dieser Fragmente wieder zusammen.)) Für Marker-Gen-Studien ist Illumina MiSeq mit 1.5e10 Basen und ca. 300 Basen-langen Sequenzen das aktuelle Standardverfahren.
[Quince *et al.* (2017)]

Bridge Amplification: Viele Forward- und Reverse-Primer sind an einen Chip-förmigen Träger gebunden, der sich in einer Durchflusszelle befindet. Die DNA-Fragmente lagern sich an jeweils einen Primer an. Freie DNA-Stücke werden weggespült. Die meisten Primer bleiben zunächst frei. DNA-Polymerase verlängert den angelagerten Primer, so dass ein DNA-Doppelstrang vorliegt. Die Stränge werden getrennt. Jeder Strang hat die Möglichkeit, sich zu

biegen und mit der Primersequenz an seinem Ende an einen der Primer, die an den Chip gebunden sind, zu hybridisieren. (Das DNA-Fragment verbindet die beiden Chip-gebundenen Primer wie eine Brücke.) An diesem startet eine weitere DNA-Synthese. Der Prozess wird wiederholt und führt zu räumlichen Clustern identischer DNA-Einzelstränge.
[Metzker (2010), Sinha *et al.* (2017)]

Sequenzierung: Es werden Primer für eine Richtung hinzugefügt. Es werden alle vier Arten von dNTPs hinzugefügt. Jede Art von dNTP ist an einen fluoreszierenden Farbstoff gebunden und besitzt zusätzlich eine Blockierung, die die Verlängerung der Kette nach seinem Einbau verhindert. Nachdem die nicht eingebauten dNTPs weggespült wurden, wird der Farbstoff mit einem Lichtimpuls (Laser) angeregt und anschließend das emmitierte Licht gemessen. Der Farbton gibt die eingebaute Base an. Der Farbstoff wird abgespalten und weggespült. Die Blockierung wird entfernt. Der Prozess wird wiederholt, bis die Sequenz bekannt ist. Der eigentliche Sequenzierungsprozess kann mit dem zweiten Primer wiederholt werden, wodurch hohe Signalqualität für mehr Basenpaare erreicht wird.
[Metzker (2010)]

Sequenzierungsfehler: Substitutionen sind der häufigste Fehlertyp, wobei nach einem G die Fehlerrate am höchsten ist. Die Unterrepräsentation von AT und GC reichen Regionen ist auffällig bei Illumina, was wahrscheinlich auf einen Amplifikations-Bias zurückzuführen ist.
[Metzker (2010)]
Das ExAmp-Verfahren führte zusätzliche Fehler ein.

ExAmp: Illumina ersetzte 2015 die Bridge Amplification in den neueren HiSeq Geräten durch ein ExAmp genanntes Verfahren. Die Einzelheiten zu ExAmp sind nicht bekannt. ExAmp enthält keine Binde-und-Wasch-Schritte vor der Cluster-Erzeugung. Durch das Einbringen von kleinen Vertiefungen, sogenannten Nanowells, in der Durchflusszelle, konnte die Clusterdichte erhöht und die Kosten gesenkt werden. Die Sequenzierungsfehler wurden reduziert und die Sequenzlängen erhöht. Allerdings wurden die Fehler gravierender. Bei allen NGS-Verfahren werden in vielen Experimenten mehrere Proben gleichzeitig sequenziert um Kosten zu sparen. Um sie später auseinanderhalten zu können, werden sie zuvor mit einer Barcodesequenz versehen. ExAmp führt Kreuz-Kontamination: 5% bis 10% der Sequenzen enthalten die falschen Barcodes und werden daher anschließend den falschen Proben zugeordnet.
[Sinha *et al.* (2017)]

C.1.1.4. Pacific Biosciences – PacBio (Third Generation Sequencing)

Pacific Biosciences kann vollständige oder fast vollständige isolierte mikrobielle Genome mit geringer Fehlerrate sequenzieren, wenn das Genom mindestens 30 mal sequenziert wird.

Das Verfahren produziert inzwischen 1e10 Basen insgesamt pro Durchlauf und wird damit zunehmend auch für Metagenomstudien interessant.
[Quince *et al.* (2017)]

Sequenzierung: Polymerase Moleküle sind auf einem Träger fixiert. Das DNA-Fragment mit Primer lagert sich daran an. Hierdurch ist die Position des enzymatisch-aktiven Teils der Polymerase und somit die Position, an der das Signal zu erwarten ist, bekannt. Die Form des Trägers an dieser Stelle und die Art des zur Anregung verwendeten Lichtreizes tragen maßgeblich dazu bei, dass das Signal eines einzelnen Moleküls gemessen werden kann. Im Gegensatz zu den vorherigen Verfahren ist keine Amplifikation nötig. Die Synthese eines einzelnen DNA-Strangs kann in Echtzeit beobachtet werden. Als Signal dienen mit speziellen Farbstoffgruppen markierte Nukleotide. Die Abspaltung dieser Gruppe erzeugt einen entsprechenden Lichtimpuls. Mit diesem Verfahren können längere Sequenzen bestimmt werden.
[Metzker (2010)]

Sequenzierungsfehler: Die Korrektheit der Auslesung lag ursprünglich bei 83%. Fehler entstehen unter anderem, wenn die DNA-Synthese zu schnell abläuft oder wenn ein Nukleotid das aktive Zentrum der Polymerase eindringt aber nicht in die DNA eingebaut wird. Die meisten Fehler scheinen gleichverteilte Zufallsereignisse zu sein. Daher kann eine Korrektheit von über 99% erreicht werden, wenn ein DNA-Fragment mehrfach sequenziert wird. ((Da die Sequenz aber nur über die Polymerase an den Träger gebunden ist, ist das möglicherweise für Mikrobiomanalysen nicht trivial.))
[Metzker (2010)]

C.2. Next Generation Sequencing Experimente

C.2.1. Marker-Gen-Analysen

Marker-Gen-Analysen eignen sich dafür einen Überblick mit geringer phylogenetischer Auflösung zu erhalten. Sofern genügend Geld zur Verfügung steht empfiehlt es sich anschließend Metagenom- oder Metatranskriptomanalysen durchzuführen. Marker-Gen-Analysen basieren auf Primern für spezielle Regionen von interessierenden Genen, die dann genutzt werden um die Phylogenie einer Probe zu bestimmen und auszuzählen. Die Vermehrung zur anschließenden Sequenzierung, sowie die Sequenzierung selbst, von Marker-Genen wie 16S rRNA für Protisten und ITS für Pilze sind gut etabliert, schnell und preiswert. Für andere Eukaryoten wird meist 18S verwendet. Da DNA von Organismen sich in der Zielsequenz für die Primer unterscheiden, lagern sich die verwendeten Primer unterschiedlich gut an und verursachen folglich Verzerrungen der Häufigkeiten während der PCR. Die Auswahl mehrerer geeigneter Primer kann dem entgegenwirken, erfordert jedoch a priori Wissen über die mikrobiellen Gemeinschaften. Häufig sind selbst gut optimierte Primer-Sets nur bis auf Genus-Ebene wirklich wirksam. Marker-Gen-Analysen stellen die geringsten Anforderungen an

Proben und Studien Design. Sie funktionieren noch halbwegs gut mit Stichproben, die durch Wirt-DNA verunreinigt sind, und Proben, die wenig Zellen enthalten. Allerdings sind Proben mit wenig Zellen besonders anfällig für Verzerrungen durch PCR-bedingte überrepräsentation von konterminierenden Mikroben. Phylogenetische Bäume sind gewöhnlicherweise sehr ungenau, wenn sie direkt aus Marker-Gen-Sequenzen erstellt werden, so dass stattdessen eine Einbettung in einen Referenzbaum empfohlen wird. Außerdem besitzen einige bakterielle Familien, die sich genomisch, phänotypisch und funktionell deutlich unterscheiden, sehr ähnliche 16S Regionen.
[Knight *et al.* (2018)]

Ideale Marker: Bei ausreichend langer Zeit, zwischen zwei Sequenzen, bis zum MRCA ((Most Recent Common Ancestor - d.h. die jüngste Sequenz, aus der sich beide mit der Zeit entwickelt haben)), können mehrere Punktmutationen an derselben Sequenzposition auftreten. Phylogenetische Information geht dabei verloren und kann zu falschen Schlüssen verleiten. Ein guter phylogenetischer Marker sollte deshalb sehr konservierte Regionen als Targets für universelle Primer und zur Bestimmung von phylogenetischen Beziehungen zwischen entfernten Organismen haben, aber auch variablere Regionen für phylogenetische Beziehung auf Familien- und Gattungsebene.
[Case *et al.* (2007)]

C.2.1.1. Internal Transcribed Spacer Barcodes

Der offizielle Barcode-Marker für Pilze ist ihr internal transcribed spacer (ITS) Lokus. ITS ist grundlegend in vielen phylogenetischen und ökologischen Studien und für die Identifizierung von Pathogenen. Der ITS-Lokus besteht aus den zwei nicht-kodierenden Regionen ITS1 und ITS2, die die, für die 5.8S rRNA Untereinheit kodierende, Sequenz umgeben. Die 5.8S Region ist sehr konserviert, selbst über taxonomische Königreiche hinweg. Die beiden nicht-kodierenden Regionen sind sehr variabel. Die ITS-Region enthält sehr viele Indels. Bei weitem die meisten sind nur eine Base lang. Nur wenige in dieser Region sind länger 20 Basen. Die ITS-Regionen der nuklearen ribosomalen DNA sind die häufigst verwendeten phylogenetischen Marker bei Pilzen. ((Mit nuklear ist gemeint, dass sie im Zellkern liegt, im Gegensatz zu DNA in den Mitochondrien.)) Die ITS-Region liegt in vielen Kopien vor und ist daher selbst in schlecht erhaltengeblieben Proben untersuchbar. Das ist ein Grund für die Beliebtheit der ITS. Primer für die Region sind gut getestet. Die Länge von 400-800 bp der ITS eignet sich ungefähr für Sequenzierungen und kann gut zwischen verschiedenen Pilzarten unterscheiden. Bioinformatische Werkzeuge für Pilze basieren häufig auf ITS-Sequenzen. Besonders aufgrund von Indels ist ein multiples Alignment von entfernt verwandten ITS-Sequenzen problematisch. Fehlalignments begünstigen anschließend phylogenetische Fehlzuordnungen. Die Substitutionsraten in Bereichen außerhalb der indelreichen Regionen, sind allerdings so groß, dass ähnliche Sequenzabschnitte unabhängig voneinander in mehreren Abstammungslinien entstehen, was auch ohne Indels zu Problemen bei der Konstruktion

eines phylogenetischen Baums führt. Die besonders vielen Indels und die fehlende phylogenetische Auflösung jenseits der Artenebene, sind die beiden Hauptkritikpunkte an ITS. Daher wird davon abgeraten aus ITS phylogenetische Information zu berechnen und sie nur für Zuordnung zu Arten zu verwenden. Durch geschicktere Nutzung der Indels lässt sich die Auflösung verbessern, da sie etwas konservativer sind als Substitutionen. Hierdurch wird die phylogenetische Information des ITS-Lokus auch auf Gattungsebene verwendbar und auch noch darüber. Es ist allerdings zu Berücksichtigen, dass phylogenetische Konstruktionen auch davon abhängen, wie gut die gemessenen Mikroben für Aussagen über gemeinsame Vorfahren geeignet sind. Wenn mehr unterschiedliche Mikroben gemessen werden, erhöht sich tendenziell die relative Anzahl ähnlicher Mikroben und damit die Information über innere Knoten im phylogenetischen Baum, wodurch dieser leichter konstruierbar wird.
[Nagy *et al.* (2012)]

C.2.1.2. DNA-Barcode für Pilze

Zur Identifikation möglichst alle Arten der taxonomischen Königreiche der eukaryotischen Domäne, wurden DNA-Barcodes bestehen aus standardisierten 500 bis 800 Basenpaar-langen Sequenzen vorgeschlagen. Die benötigten Primer sollen für die breiteste Gruppe taxonomische Gruppe passen. Vom "Consortium for the Barcode of Life" wurde die ITS-Region als DNA-Barcode für Pilze aus 6 Kandidaten-Regionen ausgewählt. Unter den ribosomalen Regionen führte ITS zur besten Identifizierbarkeit über die weiteste Menge an Pilzen. Die nukleare ribosomale kleine Untereinheit zeigte die schlechteste Auflösung auf Artenebene. Protein-kodierende Gene führten zwar häufig zu besser Identifizierbarkeit als ribosomale Marker, waren aber schwer zu amplifizieren und sequenzieren.
[Schoch *et al.* (2012)]

Mitochondriale Marker: Die Region der mitochondrialen Cytochrome C Oxidase Untereinheit 1 (CO1), die als Barcode für Tiere verwendet wird, wurde nicht gewählt, weil sie in Pilzen kompliziert zu amplifizieren ist, oft lange Introns enthält und manchmal nicht genügend variabel ist. Innerhalb derselben Art existiert ein Intron manchmal und manchmal nicht, manchmal in mehreren Kopien unterschiedlicher Länge und mit identischer Sequenz in anderen Arten. Da die meisten Pilze mikroskopisch und oft unkultivierbar sind, ist ein guter universeller Primer nötig, was bei CO1 anscheinend nicht möglich ist. Wenn in einer entsprechenden Studie dieselbe Barcode-Region für alle Königreiche (in der Domäne Eukarya) verwendet werden soll, soll allerdings CO1 verwendet und für Pflanzen ein zusätzlicher Marker hinzugezogen werden. Einige Pilzarten haben keine Mitochondrien.
[Schoch *et al.* (2012)]

Protein-kodierende Marker: Protein-kodierende Gene werden häufig in phylogenetischen Pilz-Studien verwendet. Aus ihnen wird viel mehr Information über phylogenetische Verwandtschaften gezogen als bei rRNA-Genen. Die benötigten Primer passen jedoch nur

auf enge taxonomische Gruppen. Allerdings existiert ein Protein-kodierendes Gen, die größte Untereinheit der RNA polymerase II (RPB1), das als Barcode für Pilze hätte verwendet werden können, es gelang jedoch nicht, es für alle Arten der Familie *Glomeraceae* zu amplifizieren.
[Schoch *et al.* (2012)]

C.2.2. Metagenomics

Bei "Whole Metagenome Analysen" werden alle mikrobiellen Genome in einer Probe sequenziert. Dies liefert höhere Auflösung als der Marker-Gen-Ansatz, ist allerdings teurer und aufwendiger in der Probenpräparation, Sequenzierung und Auswertung. Jede DNA in der Probe ist der Methode zugänglich. Das schließt Viren-DNA und eukaryotische DNA ein. Bei geeigneter Sequenztiefe und taxonomischer Auflösung ist eine ungenaue virtuelle Rekonstruktion ("assembly") weitgehend vollständiger mikrobieller Genome aus kurzen Reads möglich. Sofern Funktionen von Genen bekannt sind, lassen sich diese bei der Auswertung berücksichtigen, bzw. auch die Auswertung auf diese fokussieren. Verzerrungen entstehen durch die Präparation der Proben, die virtuelle Rekonstruktion und durch die Referenzdatenbank zur Funktionszuschreibung. Diese Verzerrungen sind aufgrund mangelnder Erfahrung momentan weniger gut verstanden als die beim Marker-Gen-Ansatz.
[Knight *et al.* (2018)]

Pilz-Metagenomik: In Menschen kommen Bakterien wesentlich häufiger vor als Pilze. Pilze machen weniger als 0.1Daher sind hier metagenomische Analysen für Pilz-Biome besonders teuer und schwer zusammensetzbar.
[Tiew *et al.* (2020)]

C.2.2.1. Gen-zentrische Verfahren

Gen-zentrische Ansätze, die Gene identifizieren und funktionellen Kategorien zuweisen, erleichtern funktionelle Vergleiche von Umgebungen. Die Gene werden dabei keinen Organismen zugeordnet. Solche Analysen sind momentan aufgrund der hohen Anzahl an Genen ohne bekannte Funktion und unbekannter Wechselwirkungen zwischen Genen nur eingeschränkt möglich.
[Wilmes *et al.* (2008)]

C.2.3. Metatranskriptom

Die Gesamtheit aller Gen-Transkripte (mRNA) einer Organismengemeinschaft wird Metatranskriptom genannt. Im Metatranscriptomics-Ansatz wird diese RNA sequenziert. Die Information entspricht der aktiven funktionellen Produktion des Mikrobioms. Im Gegensatz

zum obigen Ansatz werden bei dieser Methode nur lebende Organismen, und von diesen nur die aktiven Funktionen, bei der Sequenzierung berücksichtigt. Da RNA instabiler als DNA ist, ist die Lagerung aufwendiger.
[Knight *et al.* (2018)]

C.2.4. Technik und Versuchswesen

Labortechnische Unterschiede in experimentellen Methoden können die Ergebnisse sehr beeinflussen. Dies gilt auf allen Ebenen von der DNA-Extraktion bis zur Sequenzierung. Daher müssen diese für den gesamten Versuch konstant gehalten werden. Mikrobiomdaten sinnvoll zu analysieren, für die unterschiedliche Methoden verwendet wurden, ist ein ungelöstes Problem. In Longitudinalstudien müssen mehrere Baseline-Proben genommen werden. Leerproben sind essenziell um Verunreinigungen festzustellen. Dies gilt für die Probenentnahme, DNA-Extraktion, PCR und das Sequenzieren. Proben sollten bei $-80°C$ gelagert und transportiert werden. Falls dies nicht möglich ist, kann beispielsweise eine 95%-ige Alkohol-Lösung verwendet werden. Die Verwendung von Referenzproben mit bekannten Zusammensetzung in jedem Sequenzierungsdurchlauf wird empfohlen um Analysen zu standardisieren.

[Knight *et al.* (2018)]

C.2.5. Verzerrungen in Mikrobiomanalysen

C.2.5.1. Unterschiedliche Anzahlen des Marker-Gens

Marker-Gene können mehrfach im Genom vorliegen, wobei sich die Anzahl sich zwischen Arten unterscheidet, wodurch der Eindruck entsteht, dass Arten mit mehr Kopien auch häufiger in der Probe vorkommen. Insbesondere haben Bakterien, die auf eine bestimmte Umwelt spezialisiert sind, tendenziell weniger rRNA-Gen Kopien als solche, die in vielen Umgebungen wettbewerbsfähig sind.
[Case *et al.* (2007)]

C.2.5.2. Intragenomische Heterogenität

Kopien von 16S können innerhalb desselben Organismus Sequenzunterschiede aufweisen. Das wird "intragenomische Heterogenität" genannt. Die gesamte 16S Sequenz zeigt diese Heterogenität, jedoch sind einige Strukturen der rRNA Sekundärstruktur besonders betroffen. Alle 16S Bereiche, die gewöhnlich in der mikrobiellen Ökologie verwendet werden, besitzen solche. Intragenomische Heterogenität in mehreren Kopien sind auch bei Eurkaryoten in 18S bekannt (z.B *Plasmodium berghei*). Intragenomische Heterogenität beeinflusst die OTU-Bildungen, phylogonetische Auflösung und Gene-Tree-Topologie (hauptsächlich auf Arten-Ebene und darunter). Sie hat am wahrscheinlichsten Einfluss auf fein-skalierte Phylogenetik,

da dort die intergenomische Variabilität mit der intragenomischen am leichtesten verwechselt werden kann. Bei 16S wurde intragenomische Heterogenität von bis zu 11.6% zwischen den Positionen beobachtet. Im Vergleich dazu wird als OTU Definition gewöhnlich zwischen 97% und 98% Sequenzübereinstimmung angesetzt.
[Case *et al.* (2007)]

C.2.5.3. Heterogenität innerhalb einer Art

Marker-Gen-Sequenzen brauchen nicht bijektiv zu Arten zuordnenbar zu sein. Beispielsweise existieren Bakterien, bei denen das 16S Gen nur einmal vorkommt, aber Sequenzdiversität innerhalb der Art anzutreffen ist (z.B *Vibrio splendidus*).
[Case *et al.* (2007)]

C.2.5.4. Primerdesign

Wie häufig eine Art – und ob sie überhaupt – sequenziert wird, hängt unter anderem davon ab, wie gut die verwendete Primersequnenz zu ihr passt. Sogenannte universelle Primer enthalten degenerierte Basen an einigen Stellen, von denen bekannt ist, dass Bakterien mit abweichenden Nukleotiden existieren. ((Degenerierte Basen behindern keine Anlagerung von Basen.)) Aber selbst einige weitverbreitete universelle 16S Primer führen zu deutlicher Unterrepräsentation von Sequenzen einiger Arten. Es empfiehlt sich daher die Sequenzierung der Proben mit verschiedenen Kombinationen von Extraktionsmethoden und Primern zu wiederholen.
[Walker *et al.* (2015)]

C.2.5.5. DNA-Extraktion

Mechanische DNA-Extraktionsverfahren werden für besser gehalten als chemische. Sie können jedoch zu kürzeren DNA-Fragmenten und weniger gemessenen Sequenzen führen, was wiederum dazu führt, dass viele seltene OTUs fehlen und die Diversität der Probe unterschätzt wird.
[Quince *et al.* (2017), Walker *et al.* (2015)] Die Anzahl und Beschaffenheit der Zellwände beeinflusst die Extraktion der DNA und keine Extraktionsmethode funktioniert gleich gut für alle Typen. Mechanische Verfahren wie Bead-Beating erhöhen die Repräsentativität, behindern allerdings durch kürzere DNA-Fragmente solche Sequenzierverfahren, die lange Sequenzen lesen können.
[Quince *et al.* (2017)]

C.2.5.6. Kontamination

Kontaminationen verfälschen Ergebnisse und können bereits in Reagenzen vorhanden sein, die zur Behandlung der Proben gekauft wurden. Das wirkt sich besonders negativ auf Proben mit geringer Anzahl an Bakterien aus, wie es bei Haut-Swabs der Fall ist. Kontroll-DNA, die üblicherweise zu Protokollen für Illumina-basierte Sequenzierung gehören, kann zu Kontaminationen führen. ((Illumina ist der Hersteller der aktuell meist verwendeten Sequenzierautomaten.)) Ebenso Wirts-DNA und DNA der am Versuch beteiligten Personen.
[Quince *et al.* (2017)]

Index Hopping: Künstliche DNA-Barcodes werden unter anderem zur Zuordnung von Proben verwendet. Index Hopping bezeichnet den Einbau eines Barcodes in Sequenzen aus einer anderen Probe. Dies führt zu Fehlzuordnung. Bei Illumina Plattformen können Überbleibsel vorheriger oder aktueller Benutzungen des Sequenzers die Ergebnisse verfälschen. Illumina Plotformen unterscheiden sich hauptsächlich in ihrer Sequenzierleistung und produzierenden Sequenzlängen, jedoch führte die Einführung des Verfahrens "ExAmp" zu hohen Raten von Index Hopping.
[Quince *et al.* (2017)]

C.2.5.7. PCR

Beim Amplifizieren von DNA-Fragmenten mittels PCR entstehen Mutationen. Außerdem werden Fragmente, die viele AT ((d.h. in den Sequenz folgt ein T auf ein A)) oder GC aufweisen werden schlechter amplifiziert und sind daher unterrepräsentiert unter der vervielfältigten Molekülen Metzker (2010). PCR-Biases können durch PCR-freie Methoden wie TruSeq vermieden werden Quince *et al.* (2017). Bei der PCR entstehen leicht Chimären ((Fragmente, die sich aus unterschiedlichen Sequenzen zusammensetzen)) Parks *et al.* (2018).

C.2.5.8. Sample-Handhabung

Die zu untersuchenden Proben enthalten Lebewesen. Die Lagerung und der Transport der Proben kann zu Verzerrungen führen.
[Case *et al.* (2007)]

C.2.5.9. Metagenomanalysen

((Metagenomanalysen untersuchen die Genome einer Probe direkt, ohne zuvor Marker-Gene zu vervielfältigen. Die DNA wird momentan allerdings weiterhin in Fragmente zerlegt.)) Zur Zeit muss in Metagenomanalysen ein Kompromis eingegangen werden. Wenn der Fokus darauf liegt auch seltene Genome zu finden, nehmen Sequenzierungsfehler für Genome mit

langen repetitiven Sequenzabschnitten zu.
[Quince *et al.* (2017)]

Ähnliche Organismen: Eine Probe kann verschiedene Stämme derselben bakteriellen Art enthalten. Dies kann Probleme beim virtuellen Zusammensetzen der Genome aus DNA-Fragmenten verursachen und führt momentan häufig zum Abbruch. ((Bei zwei Fragmenten, mit sich teilweise unterscheidenden Basenfolgen in ihrer Mitte, ist nicht klar, welches der beiden verwendet werden sollte, da beide Puzzlestücke gleich gut an die bisherige Sequenz passen. Da anschließend wieder identische Fragmente zu erwarten sind und sich das Problem auch wiederholen kann, wächst die Anzahl möglicher Sequenzen kombinatiorisch.)) Nach dem Zusammensetzen bleiben Fragmente übrig, die nicht zugeordnet wurden.
[Quince *et al.* (2017)]

Referenzgenome: Alle aktuellen metagenomischen Bioinformatik-Tools basieren auf verfügbaren Genomen, wodurch die Auswertungen verfälscht werden, da die Menge bekannter vollständiger Genome nicht-repräsentativ ist. Solche von Modellorganismen, Pathogenen und leicht kultivierbaren Organismen sind deutlich überrepräsentiert.
[Quince *et al.* (2017)]

Verifizierte Information: Die Zuweisung von Funktionen an Gene geschieht mittels Datenbankabgleichen. Sie ist auf bekannte, verifizierte Zusammenhänge zwischen Sequenz und Funktion angewiesen. Davon gibt es nur wenige und neue kommen nur langsam aus aufwendigen Experimenten hinzu. Boden- und Wasser-Mikrobiome sind aufgrund ihrer großen Diversität und des hohen Anteils uncharakterisierter Taxa besonders betroffen. Die Repräsentativität steht auch hier in Frage.
[Quince *et al.* (2017)]

Reste toter Lebewesen: DNA ist ein langlebiges Molekül. Es bleibt nach dem Zelltod und auch außerhalb der Zelle bestehen. ((Diese können für viele Fragestellungen als eine Art natürliche Kontamination betrachtet werden. Sie sind in Marker-Gen-Studien weniger problematisch, weil dort die Marker-Sequenz intakt geblieben sein muss, während in Metagenomanalysen jede Sequenz stören kann.)) Sequenzierungen repräsentieren daher nicht zwangsläufig die lebende Mikrobengemeinschaft. In Marker-Gen-Analysen lassen sich bereits wenige Tage nach einer Schädlingsbekämpfung Unterschiede nachweisen [Poret-Peterson *et al.* (2019)]. Mit chemische Methoden kann die freie DNA und DNA in beschädigten Zellen und dieses Problem beseitigt werden. Alternativ lässt sich das Metatranskriptom ((d.h. alle mRNA)) untersuchen, das zwangsläufig aus lebenden Organismen stammt.
[Quince *et al.* (2017)]

Phylogenetische Auflösung: Momentan sind metagenomische Analysen noch nicht genau genug um zwischen Bakterienstämmen zu unterscheiden. Für gute Populationsgenetik und mikrobielle Ökologie wäre dies nötig.
[Quince *et al.* (2017)]

Entkopplung von Phylogenetik und Funktion: Lateraler Gentransfer hat bei Prokaryoten die Verwandtschaftsbeziehung von der genetischen Information teilweise entkoppelt. Insbesondere gilt das für Gene, deren Funktionen Selektionsdruck unterliegen. Zusammenhänge zwischen Bakterienverwandtschaft und Ökosystem lassen sich dann schwerer untersuchen. ((Selektionsdruck ist örtlich bestimmt, während Verwandtschaftsbeziehungen zwar an einem Ort entstehen aber auseinander migrieren.)) Bei der Analyse von einfachen ökologischen Funktionen wird daher auch auf die Sequenzen der entsprechenden funktionellen Gene zurück=gegriffen – ohne diese erst Organismen zuzuordnen. Housekeeping-Funktionen ((d.h. Funktionen die unabhängig von Umwelteinflüssen exprimiert werden)) machen lateralen Gentransfer der betreffenden Sequenz unwahrscheinlicher.
[Case *et al.* (2007)]

C.3. Next Generation Sequencing Analyse

Begriffe: Als Read wird die ausgelesene Sequence der Basenpaare eines DNA-Fragments bezeichnet. Sequenztiefe ist die Anzahl der Reads pro Probe.
[Knight *et al.* (2018)]

Standards Die Standards und Methoden für Mikrobiomanalysen entwickeln sich schnell. Es wird inzwischen empfohlen sogenannte "exakte Sequenzen" anstatt OTUs für Analysen zu verwenden. Änderungen in metagenomischer Taxonomie in den nächsten Jahren, werden ebenfalls vermutet.
[Knight *et al.* (2018)]

C.3.1. Behandlung von Sequenzierungsfehlern

Die Fehlerwahrscheinlichkeit von Illumina Sequenzierung liegt bei ca. 0.1% pro Nukleotid. Sequenzierungsfehler täuschen einen Großteil der Diversität einer Probe vor. Die Reads werden üblicherweise zu OTUs (97% Übereinstimmung) geclustert um dies zu reduzieren. Dabei geht ungünstigerweise reale Diversität verloren. In neueren Verfahren wie DADA2 [Callahan *et al.* (2016)] und Deblur [Amir *et al.* (2017)] bleibt diese besser erhalten. Deblur ist ein Filtermechanismus, der häufigeren Reads eine höhere Wahrscheinlichkeit zuschreibt korrekt zu sein. Er iteriert vom häufigsten zum seltensten Read. Die Häufigkeiten von Reads, die weniger häufig sind als der aktuelle Read, werden reduziert, wobei die Reduktion von der Anzahl der

Sequenzunterschiede zum (und der Häufigkeit des) aktuellen Read abhängt. Sinkt die Häufigkeit eines Reads auf 0, wird dieser entfernt. Häufigkeiten brauchen nicht erhöht zu werden, da es sich im Grunde um Relativdaten handelt. Die Reads, die am Ende des Vorgangs noch vorhanden sind, werden "exakte Sequenzen" oder auch "sub-OTUs" genannt. Das Verfahren, insbesondere die Wahl seiner Parameter, ist auf Illumina-Sequenzierungen optimiert.
[Amir *et al.* (2017),Knight *et al.* (2018)]

C.3.2. Taxonomische Zuweisung

Üblicherweise wird die taxonomische Einordnung der (sub-)OTUs durch einen naive Bayes Klassifizierer durchgeführt, der trainiert wurde auf Genus-Ebene ca. 80% korrekte Zuweisungen zu erzielen. Exakte Zuordnungen über einen Abgleich mit Referenzdatenbanken führt zu weniger zuordbaren (sub-)OTUs. Unklassifizierte Sequenzen könnten aus der DNA von Chloroplasten oder Mitochondrien kommen. Das sollte überprüft und diese gegebenenfalls entfernt werden.
[Knight *et al.* (2018)]

C.3.3. Datenstruktur

Nach bioinformatischer Aufbearbeitung umfassen die Daten eine Matrix relativer Häufigkeiten (Probennummer x OTU oder Gen). Die OTUs sind idealerweise zusätzlich taxonomisch zugeordnet und einem phylogenetischen Baum angeordnet und Gene mit Funktionen versehen. Die Matrix ist hochdimensional und dünnbesetzt. Sie bestehen häufig aus tausenden von OTUs und benötigen besondere statistische Herangehensweisen. Die Daten werden üblicherweise mit Alpha- und Beta-Diversitätsmaßen untersucht. Die verwendeten phylogenetischen Größen hängen viel stärker von der Anzahl der OTUs in einer Probe ab als Größen, die nur die Häufigkeitsinformation nutzen. Phylogenetische Größen wurden nur für 16S Daten evaluiert. Bei der Auswahl der zu betrachtenden Größe sollte die biologische Interpretierbarkeit beachtet werden. Phylogenetische Größen liefern interpretierbare biologische Muster, benötigen jedoch phylogenetische Bäume. Um Diversitätsmaße sinnvoll einsetzen zu können, muss die Summe der Häufigkeiten der Sequenzen in einer Probe ("sampling effort") angeglichen werden. Diese kann sich zwischen zwei Proben um mehrere Größenordnungen unterscheiden. Rarefaction ist die aktuell beste Lösung für UniFrac.
[Knight *et al.* (2018)]
Bei Rarefaction wird aus jeder Probe eine Stichprobe gezogen, die so groß ist wie die kleinste Probe, ggf. in einer Jackknife-Schleife, um die Information aus der größeren Proben besser nutzen zu können.

C.3.4. Kompositionelle phylogenetische Auswertungs-Methode

Relativdaten werden auch "kompositionelle" Daten genannt. Die meisten statistischen Standardverfahren liefern auf Relativdaten unpassende Ergebnisse. Die Erhöhung einer Variable täuscht das Absinken aller anderen vor. Dies kann umgangen werden, wenn die Daten zuvor geeignet transformiert werden. Die isometrische Log-Ratio (ILR) Transformation eignet sich hierfür, sofern sie durchführbar ist. ((Definitionsbereich des Logarithmus.)) Die ILR kann über eine sequentielle binäre Partition (d.h. binären Baum ohne Astlängen) der Variablen konstruiert werden. Die Interpretation der Ergebnisse hängt von der Wahl der Partition ab. Wenn für Mikrobiomanalysen hierzu ein phylogenetischer Baum verwendet wird, nennt sich die Methode: phylogenetische ILR (PhILR) Transformation. Euklidische Abstände auf diesen transformierten Daten scheinen eine kompositionell-robuste Alternative zu Bray-Curtis-, Jaccard- und UniFrac-Abständen auf den Originaldaten zu sein. Auf Testdatensätzen führten PhILR-transformierte Daten als Eingabe für die Klassifikationsverfahren logistische Regression, Support Vector Machines, k-nearest Neighbors und Random Forests zu gleich guten oder besseren Ergebnissen (bzgl. Accuracy) als untransformierte oder log-transformierte Daten. PhILR bildet $D > 1$ OTUs auf ein $D - 1$ dimensionales Koordinatensystem ab. Die Koordinaten werden "Balancen" genannt. Jede Balance y_i^* gehört zu einem inneren Knoten i des Phylogenetischen Baums mit D Blättern. Jede Balance repräsentiert den Logarithmus des Quotienten der gewichteten geometrischen Mittel der relativen Häufigkeiten der beiden Teilbäume des inneren Knotens. ((Das Vorzeichen hängt willkürlichen von der Reihenfolge der beiden Teilbäume ab.)) Zur Wahl der Gewichte werden zwei Möglichkeiten vorgeschlagen:

1. Das Produkt zwischen dem geometrische Mittel der um 1 erhöhten absoluten Häufigkeiten der zu gewichtenden OTU über alle Stichproben mit der euklidischen Norm der relativen Häufigkeiten dieser OTU über alle Stichproben. Das ist eine Heuristik, die im Datensatz der Entwickler dieser Methodik gut funktioniert hat.
2. Die Quadratwurzel der Summe der beiden Kantenlängen von der OTU zu ihren beiden Kindknoten. Die Quadratwurzel war in den Daten der Entwickler besser als andere Transformationen.

Die erste Gewichtungsart berücksichtigt die hohe Anzahl an Werten nahe 0 oder gleich 0, die in Mikrobiomdaten auftreten.
[Silverman *et al.* (2017)]

Astlängen-Gewichtung: Die zweite Gewichtungsart berücksichtigt phylogenetische Astlängen. Einige Analysen könnten von der Verwendung evolutionärer Abstände zwischen Taxa profitieren, da diese in Verbindung mit Anpassungen an neue Ressourcen stehen könnten. Im menschlichen Mikrobiom könnten Balancen nahe der Blätter mit Adaptionen an Körperstellen zusammenhängen. Die Gattung Bacteroides relativ zur Gattung Prevotella unterschied zwischen Stuhlproben und anderen Körperproben. Balancen zwischen Arten derselben Gattung konnten zwischen nahgelegenen Körperstellen unterscheiden. Die Varianz einer

Balance lässt Rückschlüsse über die Auswirkungen unterschiedlicher Habitate (Probenarten) auf die betroffen Mikroben zu. Eine Varianz nahe 0 spricht hierbei für eine gleichmäßige Anpassung der Abundanzen dieser Mikroben. Im Gegensatz zur Pearson Korrelation ist dieses Maß robust gegenüber kompositionellen Effekten. Die Varianzen der Balancen waren, im Datensatz der Entwickler, größer wenn sie nahe des Wurzelknotens lagen. PhILR war in einigen Fällen nicht besser als andere ILR-Transformationen. Der Vorteil von PhILR liegt in der besseren Interpretierbarkeit.
[Silverman *et al.* (2017)]