Vergleich von Methoden für die Analyse der Mikrobiom β-Diversität

Zusammenfassung

Der menschliche Körper ist von Mikroben besiedelt. Die Gesamtheit dieser Mikroben nennt man Mikrobiom. Durch eine Störung des Mikrobioms können verschiedene Krankheiten, wie zum Beispiel eine chronisch entzündliche Darmerkrankung oder Diabetes, hervorgerufen werden. Hochdurchsatz-Sequenzierung in Verbindung mit Bioinformatik schafft die Basis für ein besseres Verständnis des Mikrobioms. Auf diesem Wissen basierend, können Interaktionen zwischen dem Mikrobiom und Krankheiten besser verstanden und analysiert werden, sowie neue therapeutische Verfahren entwickelt werden. Das Ziel dieser Arbeit besteht darin, den Einfluss von Distanzmaßen und verschiedenen Ordinationsverfahren zu analysieren. Dazu wurden verschiedene Distanzmaße und Ordinationsverfahren verglichen. Es wird die Distanz nach Bray-Curtis, Jaccard und Anderberg sowie die Euklidische, die gewichtete- und ungewichtete Unifrac Distanz berechnet. Als Ordinationsverfahren wurden die Hauptkomponentenanalyse (PCA), die Hauptkoordinatenanalyse (PCoA) und die nicht-metrische multidimensionale Skalierung (NMDS) verglichen. Die automatisierte Analyse der Mikrobengemeinschaft erfolgt mit Hilfe eines R-Skripts, welches die genannten Distanzen und Ordinationsverfahren berechnet. Durch diese automatisierte Analyse wird die Interpretation Mikrober Daten vereinfacht. Das R-Skript kann über die Kommandozeile gestartet werden und liefert ein PDF mit allen Ergebnissen. Abschließend wurden zwei Datensätze analysiert und die Ergebnisse verglichen.

Die Resultate dieser Arbeit zeigen, dass die gewichtete Unifrac Distanz die besten Ergebnisse, basierend auf der Gesamtvarianzaufklärung, liefert. Eine gute Alternative stellt die Bray-Curtis Distanz dar, die annähernd so gute Ergebnis liefert wie die gewichtete Unifrac Distanz. Die beste Ordinationsmethode stellt die Hauptkomponentenanalyse dar, da sie auf den ersten Achsen die meiste Gesamtvarianzaufklärung besitzt. Ein weiterer Vorteil der Hauptkomponentenanalyse ist, dass sie nicht nur auf eine Distanzmatrix sondern auch auf eine OTU Tabelle angewendet werden kann. Allgemein ergaben sich beim Vergleich der Ordinationsmethoden keine wesentlichen Unterschiede man erkennt jedoch unterschiedliche Ergebnisse in Abhängigkeit verschiedener Datensätze.

Schlüsselwörter

β-Diversität, Distanzmaß, Hauptkomponentenanalyse, Hauptkoordinatenanalyse, Nichtmetrische Multidimensionale Skalierung

Analysis of Microbiome β-diversity

Abstract

The human body is inhabited by microbes. The community formed by this microbes is called microbiome. Dysfunction of the human microbiota is linked to diseases such as inflammatory bowel disease or diabetes. High-Throughput Sequencing and Bioinformatics help us to understand the role of microbial diversity in habitats within the human body. This could provide new therapeutic interventions. The aim of this thesis is to analyze different distance and ordination methods. In addition diversity measures including Bray - Curtis, Jaccard, Anderberg, Euclidean, weighted and unweighted Unifrac distance were compared. Three ordination techniques were analyzed, including principal components analysis (PCA), principal coordinate analysis (PCoA) and non-metric multidimensional scaling (NMDS).

In order to analyze the microbial community an automated R-script was developed. This script helps to analyze and interpret microbiome data. For this purpose, the above mentioned different distance measures and ordination methods were implemented. The R-script can be started from the command line and provides a final PDF with all results. The program was executed with two different sets of data. Based on the output these methods were compared. Results of this thesis show, that the weighted Unifrac distance yields the best results based on explained variability. The Bray-Curtis distance represents a good alternative though; it provides nearly as good results as the weighted Unifrac distance. The best ordination method is the principal component analysis (PCA), because the first axis captures the greatest variance of the data set. Generally, no significant differences were found when comparing the ordination methods. However, different results depend on different data sets.

Keywords

β-Diversity, Distance Measurement, Principal Component Analysis, Principal Coordinates Analysis, Nonmetric Multidimensional Scaling