|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Tab. Descriptive statistics on ratios of various distances between groups, to various distances within groups | | | | | |
| Method | Average | Std. Dev | Median | Min | Max |
| Manhattan | 9.5665 | 6.1651 | 9.2735 | 3.6321 | 16.0765 |
| Euclidean | 7.4978 | 4.1397 | 7.2055 | 3.3973 | 12.1760 |
| Canberra | 13.8394 | 9.6327 | 13.7105 | 4.7623 | 23.1759 |
| Clark | 12.1014 | 8.3447 | 12.0713 | 4.2785 | 20.0024 |
| Bray–Curtis | 9.6014 | 6.2237 | 9.3290 | 3.6229 | 16.1167 |
| Kulczynski | 9.6052 | 6.2248 | 9.3321 | 3.6252 | 16.1227 |
| Jaccard | 6.4699 | 3.9501 | 6.3592 | 2.6978 | 10.4586 |
| Gower | 8.7405 | 6.1334 | 8.5342 | 2.9049 | 14.9720 |
| Alternative Gower | 6.3777 | 3.4508 | 6.2136 | 3.0084 | 10.0681 |
| Horn–Morisita | 84.6013 | 78.7938 | 80.0459 | 10.4910 | 167.5428 |

To quantify the dissimilarity between the biofilm samples and the sludge samples, we employed multiple distance measures as summarized in **Tab.**, including Manhattan, Euclidean, Canberra, Clark, Bray-Curtis, Kulczynski, Jaccard, Gower, Alternative Gower, and Horn-Morisita distances. The ratios of various distances between groups to various distances within groups are calculated, to verify the hypothesis that samples from different categories (biofilm or sludge) should be significantly dissimilar, while samples belong to the same categories (biofilm or sludge) should be alike. In such cases, the derived ratios should all be larger than one. From **Tab.**, it is demonstrated that the differences between groups are significantly greater than differences within groups, proving that community structures are different in biofilm and sludge as measured by the listed methods.



Fig. The PCA-biplot of relative abundances at Genera level. The angles between samples represent their difference, with the scattered Genera points quantifying their contribution to the overall variation.

The PCA-biplot based on relative abundances is displayed in **Fig**. In this plot, the angles between samples (B1/B2/S1/S2) quantify similarity via their cosine values. The sharp angles between samples from the same groups, together with the almost-right angles between samples from different groups indicate that the two groups are different at community structure level via multivariate analysis. The key Genera that contributed to their differences are identified and highlighted as those far from the origin.



Fig. The DBSCAN cluster plot. The Genera are clustered according to their degree of agglomeration. The outliers are key contributors to the dissimilarity between the biofilm and sludge samples.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Tab. Key Genera that classify Sludge samples and Biofilm samples, identified by PCA-biplot and DBSCAN. | | | | |
| Genera | S1 | S2 | B1 | B2 |
| Aerolineales bacterium | 4.60% | 4.56% | 3.70% | 2.78% |
| Hassallia | 0.84% | 0.75% | 7.03% | 11.31% |
| Legionella | 1.07% | 1.14% | 12.37% | 17.46% |
| Nitrospira | 19.87% | 20.94% | 3.52% | 3.30% |
| Thauera | 12.87% | 10.96% | 1.48% | 1.45% |
| Zoogloea | 5.54% | 5.73% | 0.46% | 0.40% |

We have applied a machine learning clustering algorithm, Density-based spatial clustering of applications with noise (DBSCAN), to explore the community structure. Assuming that most Genera are shared among samples and may be trivial to flag key differences between samples, we would expect the majority of them would agglomerate into one cluster, leaving the rest as crucial identifier of samples. Accordingly, we processed the numerical ecology data and output the key different Genera as tabulated in **Tab.** . As shown above, their oscillation in relative abundances could signify the status of microbiota as sludge or biofilm.