The ISME Journal (2010) 4, 1236–1242

www.nature.com/ismej

ORIGINAL ARTICLE

Comparison of multiple metagenomes using phylogenetic networks based on ecological indices

Suparna Mitra[1], Jack A Gilbert[2], Dawn Field[3] and Daniel H Huson[1]
[1]*Center for Bioinformatics ZBIT, Tübingen University, Tübingen, Germany;* [2]*Plymouth Marine Laboratory, Prospect Place, Plymouth, UK and* [3]*NERC Centre for Ecology and Hydrology, Oxford, UK*

Second-generation sequencing technologies are fueling a vast increase in the number and scope of metagenome projects. There is a great need for the development of new methods for visualizing the relationships between multiple metagenomic data sets. To address this, a novel approach is presented that combines the use of taxonomic analysis, ecological indices and non-hierarchical clustering to provide a network representation of the relationships between different metagenome data sets. The approach is illustrated using several published data sets of different types, including metagenomes, metatranscriptomes and 16S ribosomal profiles. Application of the approach to the same data summarized at different taxonomical levels gives rise to remarkably similar networks, indicating that the analysis is very robust. Importantly, the networks provide the both visual definition and metric quantification for the non-rooted relationship between samples, combining the desirable characteristics of other tools into one.
The ISME Journal (2010) **4,** 1236–1242; doi:10.1038/ismej.2010.51; published online 29 April 2010
Subject Category: Microbial population and community ecology
Keywords: metagenomics; microbial ecology; comparative metagenomics; networks

Introduction

Metagenomics is the study of the genomic content of a sample of organisms, obtained from a common habitat or an environmental sample of microbes using sequencing. Advances in the throughput and cost-efficiency of sequencing technology are fueling a rapid growth of the number and scope of metagenomics studies, resulting in a deluge of sequences. Taxonomic analysis of such data sets has shown that only a small number of prominent taxa appear in most data sets, while the majority appear to be present only in small numbers, in what has become known as the rare biosphere (Sogin *et al.*, 2006).

There is a great need for the development of new methods for analyzing and comparing multiple metagenomic data sets, using appropriate ecological and statistical models. Explicitly, a tool that combines the visualization of relationships with a metric of distance in a single package, which includes appropriate ecological indices, without the need to fit metagenomic data to a root evolutionary dendrogramatic relationship. The two main software engineering requirements are rapid computational analysis of very large data sets and ease of use for researchers.

Correspondence: S Mitra, Center for Bioinformatics, Wilhelm Schickard Institute, Sand 14, Tübingen, BW, 72076, Germany.
E-mail: mitra@informatik.uni-tuebingen.de
Received 3 December 2009; revised 16 March 2010; accepted 20 March 2010; published online 29 April 2010

In this paper, we suggest a novel approach that combines the use of taxonomic analysis, ecological indices and non-hierarchical clustering to provide a network representation of the relationships between different metagenome data sets. The approach proceeds as follows:

First, a taxonomic profile is computed for each data set. Second, a matrix of pairwise distances is determined using one of several possible ecological indices (Legendre and Legendre, 1998). Finally, the distances are represented using an appropriate visualization technique. For reasons outlined below, we suggest to use the non-hierarchical clustering technique, neighbor-net (Bryant and Moulton, 2004).

In more detail, the first step is to produce a taxonomic profile for each given metagenomic data set. For DNA reads collected in a shotgun sequencing approach, one possibility is to use the MEGAN program (by Daniel H Huson and Stephan C Schuster (with contributions from Alexander F Auch, Daniel C Richter, Suparna Mitra & Ji Qi) Algorithms for Bioinformatics, Tuebingen University, Germany) (Huson *et al.*, 2007), which performs a taxonomic analysis of a metagenomic data set based on a BLASTX (Altschul *et al.*, 1990) comparison of a data set against an appropriate reference database such as NCBI-nr (Wheeler *et al.*, 2008). MEGAN creates taxonomic profiles at different ranks of the NCBI taxonomy, and counts how many reads are assigned to each taxon at the specified rank. The current reference databases are still largely based

on 'model organisms' and were not specifically designed as reference databases for metagenomics, thus BLAST-based analyses will be affected by the availability of good reference genomes in the database. However, the approach described in this paper is not tied to BLAST and such databases, as we show below in a study comparing 16S ribosomal RNA data.

The next step is to compute a matrix of pairwise distances from the taxonomic profiles using a suitable ecological measure. After reviewing 27 different ecological measures (listed in Legendre and Legendre, 1998), we chose six to use in this study. The simplest and most common metric measure is the 'Euclidean distance' (Equation 1), which is computed using Pythagoras' formula. The distance (D) between two metagenome samples (X, Y) can be calculated using,

$$D(X,\ Y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{1}$$

where x_i and y_i are the read counts for the i^{th} taxon of the respective metagenomic samples X and Y. It is dominated by the highly abundant taxa and its value can increase indefinitely with the number of descriptors. The Kulczynski (Equation 2) (Odum, 1950) and Bray–Curtis (Equation 3) (Bray and Curtis, 1957) distances are slightly more sophisticated measures giving by

$$D(X,\ Y) = 1 - \frac{1}{2} \times \left(\frac{\sum_{i=1}^{n} \min(x_i,\ y_i)}{\sum_{i=1}^{n} x_i} + \frac{\sum_{i=1}^{n} \min(x_i,\ y_i)}{\sum_{i=1}^{n} y_i} \right) \tag{2}$$

and

$$D(X,\ Y) = 1 - 2\frac{\sum_{i=1}^{n} \min(x_i,\ y_i)}{\sum_{i=1}^{n} (x_i + y_i)} \tag{3}$$

The χ^2 (Equation 4) (Lebart *et al.*, 1979) and Hellinger (Equation 5) (Rao, 1995) distances are two probabilistic measures that calculate the distance among sites using species abundances. They are calculated as,

$$D(X,\ Y) = \sqrt{\sum_{i=1}^{n} \frac{(\hat{x} + \hat{y})}{(x_i + y_i)} \left(\frac{x_i}{\hat{x}} - \frac{y_i}{\hat{y}} \right)^2}, \text{ with } \hat{y} = \sum_{i=1}^{n} y_i \tag{4}$$

and

$$D(X,\ Y) = \sqrt{\sum_{i=1}^{n} \left(\sqrt{\frac{x_i}{\hat{x}}} - \sqrt{\frac{y_i}{\hat{y}}} \right)^2}, \text{ with } \hat{y} = \sum_{i=1}^{n} y_i \tag{5}$$

While the 'Bray–Curtis' (Bray and Curtis, 1957) and 'Kulczynski' (Odum, 1950) measures also focus on the most abundant taxa, the 'χ^2' (Lebart *et al.*, 1979) and 'Hellinger' (Rao, 1995) distances are based on differences in the proportions of taxa between the two data sets and thus provide better representations of the taxon composition. Goodall's similarity index (Goodall, 1964, 1966) is a non-parametric measure specifically designed for determining the pairwise similarity between observations of composite multivariate data sets.

The computation of Goodall's index involves a number of steps. First, a so-called 'partial similarity measure' is calculated between each pair of species. Then for each pair of data sets, one computes the proportion of partial similarity values belonging to species i that are larger than, or equal to the partial similarity of the pair of data sets being considered. These proportions (p_i) are combined for the n species by computing the product ($\prod$) of the values relative to various species as $\prod = \prod_{i=1}^{n} p_i$. Finally the similarity (S) between two data sets (X, Y) can be obtained as the proportion of the products ($\prod$) that are larger than or equal to the product of the pair of data sets ($\prod_{\text{pair}}$) considered. The equation is giving by,

$$S(X,\ Y) = \frac{\sum_{\text{pairs}} d}{\frac{n(n-1)}{2}}, \text{ where } d = \begin{cases} 1 & \text{if } \prod \geq \prod_{\text{pair}} \\ 0 & \text{if } \prod < \prod_{\text{pair}} \end{cases} \tag{6}$$

See (Goodall, 1964, 1966; Legendre and Legendre, 1998) for further details.

By definition, Goodall's index gives more weight to differences between rare taxa than the other indices, and should therefore be particularly suitable for comparing microbial metagenomes (Sogin *et al.*, 2006).

There are two popular ways of representing distance matrices graphically. The first, widely applied in ecological studies, is to use a principal component analysis (PCA) or non-metric multidimensional scaling (NMDS) to obtain a two-dimensional layout. The second, widely used in evolutionary studies, is to use rooted trees computed by a hierarchical clustering method (Rusch *et al.*, 2007). The advantage of a tree representation is that it explicitly provides clusters of closely related data sets. However, metagenomes are not expected to evolve along a tree, rather numerous environmental factors may affect data set composition, resulting in distances that reflect incompatible signals. Although ordination methods do not suffer from this problem, they do not explicitly link data points into clusters and provide no metric against which to determine the distance between data sets. Hence, we suggest to use the neighbor-net method to compute an unrooted phylogenetic network that enjoys the advantages of both methods (Bryant and

Moulton, 2004). Such networks are not restricted to being a tree and are able to show incompatible clusters.

In this study, we apply the approach outlined above to marine metagenomes from three types of studies; a mesocosm experiment (Gilbert *et al.*, 2008), a spatially structured data set (the Global Ocean Survey) (Rusch *et al.*, 2007) and a time-series (Gilbert *et al.*, 2009). Our study suggests that the approach is robust as it produces networks that are very similar across all ranks of the NCBI taxonomy and, to a lesser extent, across different ecological indices. We further establish that the use of Goodall's index provides the best results, given that microbial communities tend to be rich in rare genes and rare taxa (Sogin *et al.*, 2006). Thus, Goodall's index may be most suitable for analyses that involve rare taxa, whereas the χ^2 and Hellinger distances can be considered when rare taxa have only a small role.

Materials and methods

All metagenomes and metatranscriptomes were aligned against the NCBI-NR database using the BLASTX tool (Altschul *et al.*, 1990). The results were imported into MEGAN (Huson *et al.*, 2007), using the 'Import from BLAST' option. To obtain taxonomic profiles, MEGAN uses the lowest common ancestor algorithm that assigns each read to the lowest common ancestor of the set of taxa that it hits in the NR database. A MEGAN project file contains all reads and all significant BLAST matches in a binary and incrementally compressed format, which is around 30% of the size of the original input files. We then performed multiple comparisons using various ecological indices and constructed networks using the neighbor-net algorithm (Bryant and Moulton, 2004), as implemented in version 4 of MEGAN.

In the first study, we compared eight Plymouth Marine Laboratory (PML)-Bergen data sets consisting of four metagenomes (DNA) and four metatranscriptomes (complementary DNA (cDNA)), and named these eight samples as follows: (1) Time1-Bag1-DNA, (2) Time1-Bag6-DNA, (3) Time2-Bag1-DNA, (4) Time2-Bag6-DNA, (5) Bag1-13May-cDNA, (6) Bag1-19May-cDNA, (7) Bag6-13May-cDNA and (8) Bag6-19May-cDNA (please refer to (Gilbert *et al.*, 2008) for details of nomenclature). All data sets were randomly re-sampled to the smallest data set size to allow inter-comparison (for example, Gilbert *et al.*, 2009). After opening all the data sets in MEGAN, the 'compare' menu item was used to generate a new document that contains a comparison of all data sets. We compared the taxonomical profiles (as MEGAN files) of these eight data sets. Then, multiple comparisons of the data sets were performed using six different ecological distance measures (Euclidean, Kulczynski (Odum, 1950), Bray–Curtis (Bray and Curtis, 1957), Hellinger (Rao, 1995), χ^2 (Lebart *et al.*, 1979) and Goodall's index (Goodall, 1964, 1966) at each of seven taxonomic ranks ('kingdom', 'phylum', 'class', 'order', 'family', 'genus' and 'species') to create a total of 42 networks (Supplementary Figures S1.1, S1.2 and S1.3). The distances were processed by the neighbor-net algorithm (Bryant and Moulton, 2004) to obtain a collection of unrooted phylogenetic networks.

In a second study, we used one random sub-sample of the Sargasso Sea data (Venter *et al.*, 2004) and one sub-sample from the Sorcerer II Global Ocean Sampling expedition data (GOS) (Rusch *et al.*, 2007) and the data and setup from the PML-Bergen study, to visualize the comparison of multiple marine metagenomes from different environments processed using different sampling and sequencing strategies. All 10 data sets were randomly re-sampled to the smallest data set size to allow inter-comparison of taxonomic abundances (for example, Gilbert *et al.*, 2009). As in the first study, we performed a multiple comparison of the 10 data sets using four of the distances (Goodall's index, Euclidean distance, Hellinger distance and χ^2 distance) at each of seven taxonomic ranks to create 28 additional networks (Supplementary Figures S2.1, S2.2), Networks obtained using the Kulczynski and Bray–Curtis distances looked very similar to the networks obtained using Euclidean distance in the previous study (Supplementary Figure S1), so we dropped the Kulczynski and Bray–Curtis distances from subsequent experiments.

In addition, multiple comparisons were performed using four of the indices considering only bacterial taxa at six taxonomic ranks, resulting in a further 24 networks (Supplementary Figure S3.1). For the Goodall's index and Euclidean distance, the numbers of sequences identified as bacterial were randomly normalized to standardize the apparent sequencing effort.

In a third study, we investigated the effect of excluding rare taxa from the taxonomical profiles. In this study, we analyzed the data at the class rank of the NCBI taxonomy. We duplicated the six metagenomes (four Bergen metagenomes, one Sargasso Sea sample and one GOS sample from the previous study) and excluded all taxa that have an arbitrarily selected abundance of $<0.025\%$ of the total community abundance from each data set. We then compared these six truncated metagenomic data sets using all six indices, resulting in six networks at the level of class taxa (Supplementary Figure S4).

In a fourth study, we analyzed all 41 samples of spatially structured GOS data. As with the previous three studies, all 41 data sets were randomly re-sampled to the smallest data set size. All data sets were 'blasted' against the NCBI-NR database and the result was imported to MEGAN. As for the Bergen samples, we computed taxonomic profiles as MEGAN files for all 41 GOS data sets. We downloaded the GOS data, from the CAMERA website (Seshadri *et al.*, 2007), we then normalized the data sets to the smallest size to allow inter-comparison of

taxonomic abundances. We performed the comparison using Goodall's index at the class rank (Supplementary Figure S5). First, we compared all the sites together (Supplementary Figure S5B) and then only the coastal and open ocean sites (Supplementary Figure S5C) to illustrate biogeographic clustering based on the assumption that the coastal sites may harbor a more diverse microbiota than the open ocean sites.

In a final study, we analyzed the correlation between 12 '16S ribosomal RNA V6 tag-pyrosequencing' data sets spanning 12 months of 2007 at a continually monitored sampling site, L4, in the Western English Channel (Gilbert *et al.*, 2009). As before, random re-sampling of these 12 samples was carried out to identical sequencing depth, to allow inter-comparison. As most operational taxonomic units (OTUs) are not present in all samples considered, we prepared an OTU abundance matrix by adding zeros in which there were no representatives for that sample.

We compared samples taken from the marine community over several months using Goodall's index in combination with neighbor-net based on all unique OTUs (Supplementary Figure S6.A), then excluding OTUs found on only one occasion (Supplementary Figure S6.B), and finally considering only the OTUs found every time (Supplementary Figure S6.C). In addition, we prepared the PCA and NMDS plots using the same OTU data for OTUs present in two or more occasions. For the PCA analysis, we used the raw data and for the NMDS calculation we used a computed Bray–Curtis matrix (Supplementary Figure S7: for a more detailed method please refer to Gilbert *et al.*, 2009).

Results and discussion

Study 1: comparison of eight marine samples from an ocean acidification study

For the PML-Bergen analysis, all six selected ecological indices produce almost identical placements of the eight samples within a neighbor network, with only minor differences in the distances between samples (see Figure 1 and Supplementary Figure S1). The placement of these PML-Bergen samples conforms to reported biological and experimental relationships (Gilbert *et al.*, 2008), with the metagenomes being well separated from the metatranscriptomes, and the samples from the peak of the induced phytoplankton bloom (Time1 or 13 May) being more separated from the samples after the collapse of the phytoplankton bloom (Time2 or 19 May) than each group is to itself. Interestingly, for the time 2 or 19 May metagenomes, the opposite is true with the differences between these being greater than their similarity to samples within the time 1 metagenomes. This is indicative of the extremely different ecology of the mesocosm samples that existed after the collapse of the bloom. This was brought about by the experimental methodology used, in which immediately after the collapse of the bloom Bag1 was re-bubbled with CO_2 and Bag6 was re-bubbled with air. This significantly altered the community composition and hence forced these samples apart (for more information refer to Gilbert *et al.*, 2008).

Study 2: comparison of multiple marine metagenomic samples from different studies

To confirm that the Bergen-PML network was robust to the inclusion of additional samples, we added two additional marine metagenomes as 'decoys'. The first was a subset of reads taken from the pooled Sargasso Sea study (Venter *et al.*, 2004) and the second was a subset of the larger GOS (Rusch *et al.*, 2007). To allow an accurate comparison, a random subset of 96 201 sequences (the size of the smallest mesocosm data set (Gilbert *et al.*, 2008)) was extracted from each study. After computing networks with four indices (Figure 2; Supplementary Figure S2), we confirmed that the eight PML-Bergen

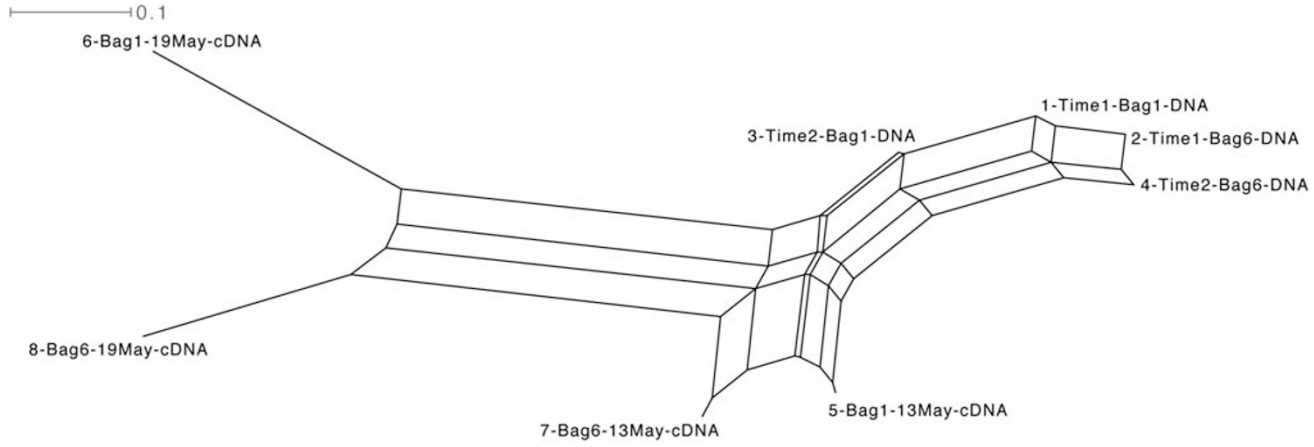

Figure 1 Network obtained using Goodall's index showing the comparison of eight PML-Bergen samples (four metagenomes and four metatranscriptomes) considering all nodes at the class rank of the NCBI taxonomy.

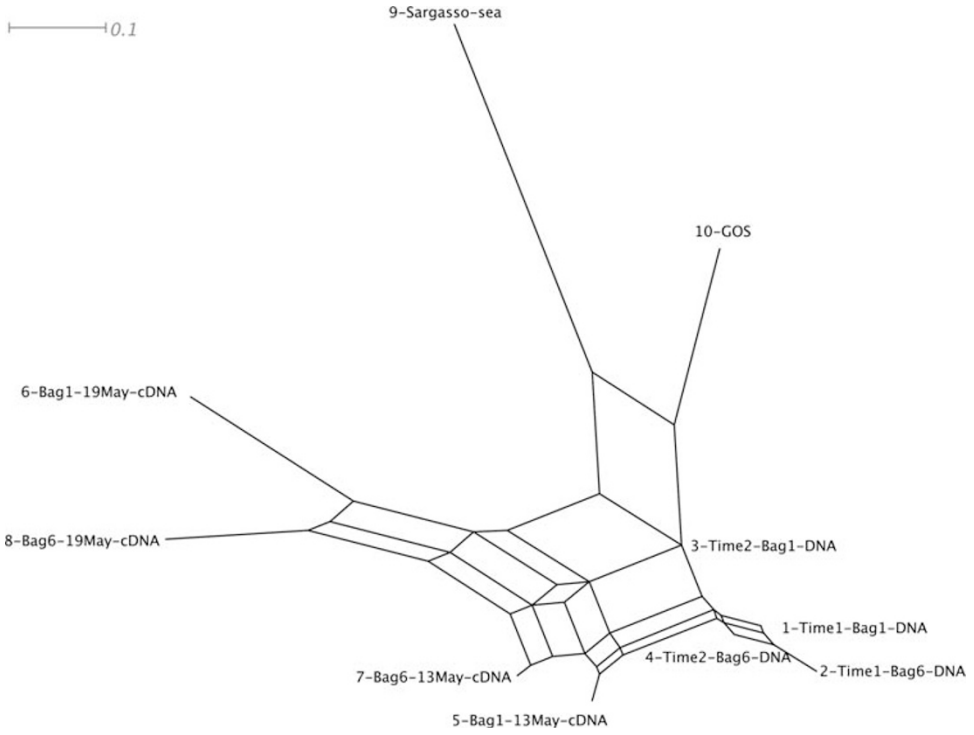

Figure 2 Network obtained using Goodall's index showing the comparison of 10 marine samples (randomly re-sampled Sargasso Sea and GOS samples together with the eight PML-Bergen samples) considering all nodes at the class rank of the NCBI taxonomy.

samples remain in their original groupings and that the two decoys are placed at a distance from them. Interestingly, there are clear differences between the networks based on the Euclidean distance, wherein the decoys are much more distantly related to the PML-Bergen samples than for the Goodall's index (Figure 2; Supplementary Figure S2), we hypothesize that this is due to the biases induced by the vast rare biosphere and the way each index handles low-abundance sequences. The networks based on the Hellinger and χ^2 distances (Supplementary Figure S2) are also similar. The GOS sample appears to cluster more closely to the PML-Bergen samples than the Sargasso Sea sample, as the GOS sample (random sub-sample of all GOS samples) is heavily enriched from coastal study sites, whereas the Sargasso Sea is an oligotrophic open ocean (Venter *et al.*, 2004).

Study 3: multiple metagenome/metatranscriptome comparisons considering only bacterial nodes

When only bacterial taxa are considered, the Sargasso Sea data set appears to be more similar to the other data sets than it does when all taxa are considered. This is because the Sargasso Sea sample contains a much smaller number of eukaryotic reads compared with the other data sets. This reflects the similar water sampling procedures (for example, filter size) for the GOS (Rusch *et al.*, 2007) and mesocosm (Gilbert *et al.*, 2008) data sets, resulting in organisms of a similar size range being analyzed; whereas the Sargasso Sea study used a different sampling procedure (Venter *et al.*, 2004), which excluded micro-eukaryotes. In this study, the networks computed using Goodall's index (Figure 3; Supplementary Figure S3) and Hellinger distance (Supplementary Figure S3) maintain a very similar layout over all ranks of the NCBI taxonomy for the 10 metagenome data sets, whereas the networks using Euclidean distance (Supplementary Figure S3) and χ^2 distance (Supplementary Figure S3) show more variability. Strikingly, unlike the first and second studies, the PML-Bergen metagenomes tend to group together by time, with time 1 (13 May) being more similar to each other than to time 2 (19 May), and *vice versa*. This suggests that the post-bloom bubbling treatment of these bags had a greater effect on the eukaryotic and archaeal communities than the bacterial communities. This is possible as a result of the bubbling-induced lysis of eukaryotic cells.

Study 4: the effect of rare taxa

To study the effect of rare taxa on such analyses, we excluded all taxa having an abundance of $<0.025\%$ from each of the six metagenomes examined above (now excluding the four metatranscriptomes). The resulting truncated data sets were then compared with the original full data sets. We observe that the

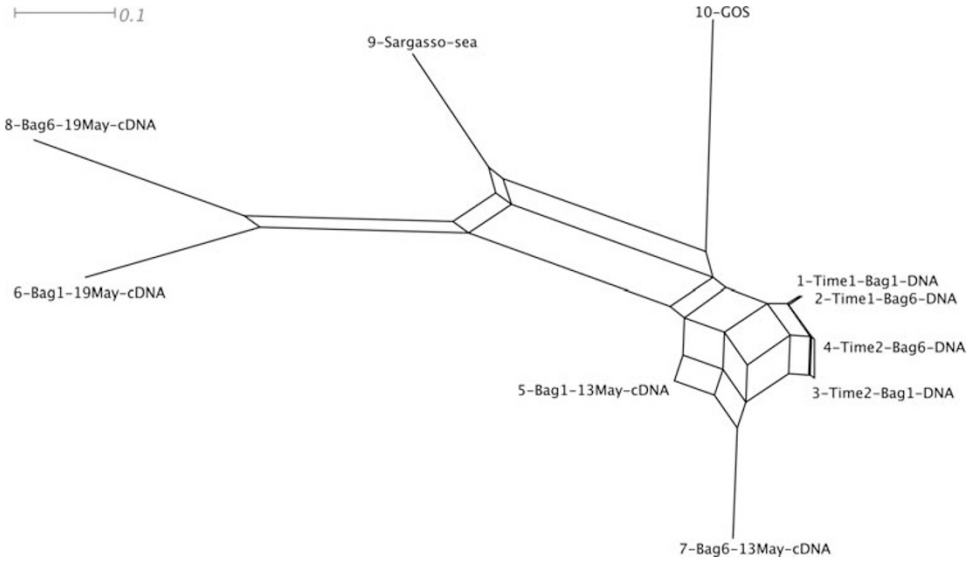

Figure 3 Network obtained using Goodall's index showing the comparison of 10 marine samples (randomly re-sampled Sargasso Sea and GOS samples and the eight PML-Bergen samples) considering only bacterial nodes at the class rank of the NCBI taxonomy.

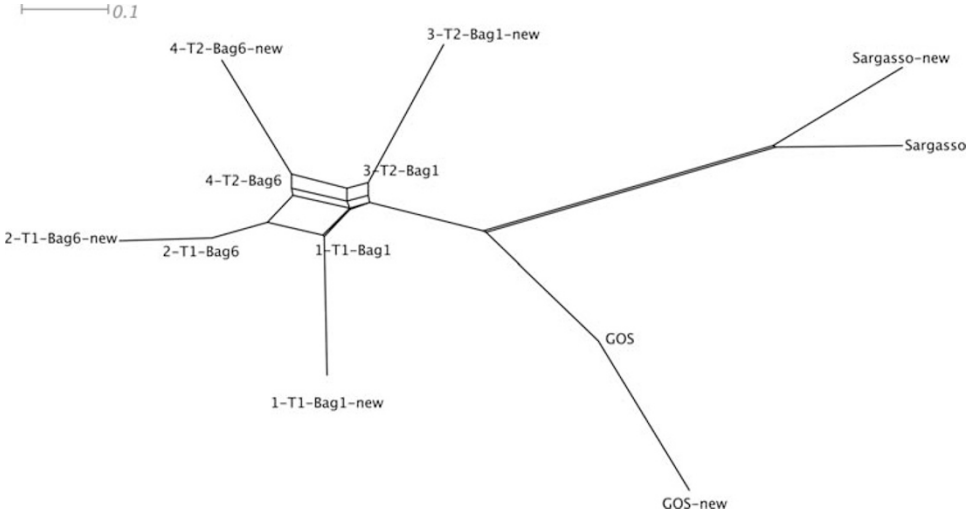

Figure 4 Comparison of six marine metagenomes (randomly re-sampled Sargasso Sea and GOS samples together with the four PML-Bergen metagenomes) with six truncated copies from which all rare taxa were excluded, analyzed at the class level of the NCBI. The displayed network is obtained using Goodall's index.

placement of the original metagenomes remains the same in all the networks computed. The networks based on the Euclidean, Kulczynski and Bray–Curtis distances are unable to distinguish between the original and truncated metagenomes, placing them at identical locations in the network (Supplementary Figure S4; left column). Networks obtained using the χ^2 and Hellinger distances place the truncated samples close to the original metagenomes, but on separate branches (Supplementary Figure S4; right column). Only the network based on Goodall's index was able to represent the correct branching within the data sets (Figure 4; Supplementary Figure S4). Interestingly, we observed that the distances between the original and the truncated data sets are roughly proportional to the percentages of community change.

Study 5: comparison of the 41 GOS data sets

We applied our approach to the geospatially structured GOS data (Rusch *et al.*, 2007) and computed two networks using Goodall's index, one considering all 41 sites and the second considering only the open ocean and coastal sites (Supplementary Figure S5). Both networks show a star-like structure, reflecting a high level of diversity in the data. Spatially related samples tend to cluster together,

with the open ocean samples showing apparently fewer sample-specific taxa than the coastal ones.

Study 6: comparison of 16S ribosomal RNA time series data from Western English Channel

To show the use of our method on 16S ribosomal RNA tag-pyrosequencing data sets, we applied it to the OTUs obtained from a continually monitored sampling site in the Western English Channel spanning February–December 2007 (Gilbert *et al.*, 2008). A comparison based on all 12 393 OTUs from this time-series data set using Goodall's index leads to a highly unresolved network (Supplementary Figure S6.A), which reflects the high abundance of rare taxa in the data across monthly samples. A more informative network can be obtained by excluding the OTUs found on only one occasion (considering 2666 OTUs, ~22%) from the analysis (Supplementary Figure S6.B). A network based only on those OTUs present in all data (71 OTUs, ~0.5%) shows similar clusters, but as a result, a proportion of the distance information is lost (Supplementary Figure S6.C). This network visually captures both the relationships between the samples and the seasonality of the data set as previously described less adequately using traditional NMDS methods (Gilbert *et al.*, 2009). This analysis highlights the robust nature of Goodall's index in marker-based metagenomic studies, as well as the importance of identifying rare taxa in these data sets.

Finally to establish the benefits of using this network representation, we prepared PCA and NMDS plot based only on those OTUs present in more than one time points (Supplementary Figure S7). Unlike the NMDS plot, the network representation (Supplementary Figure S6.B) provides a clear visualization of the distances between the different data sets, and unlike the PCA analysis it suggests possible sample groupings. An obvious direct benefit is that the network representations provide a mix of the visual sensitivity of NMDS and PCA with the quantitative nature of classical dendrograms.

Availability

A program for computing ecological indices from taxonomical profiles (called MEG2DIST) is available as open source from the website http://www-ab.informatik.uni-tuebingen.de/software/megan/meg2dist.

The code is completely integrated into version 4 of MEGAN, which is available from the website: http://www-ab.informatik.uni-tuebingen.de/software/megan.

Acknowledgements

We thank Dr Kay Nieselt for helpful discussions and Wei Wu for his assistance.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. (1990). Basic local alignment search tool. *J Mol Biol* **215**: 403–410.

Bray JR, Curtis JT. (1957). An ordination of the upland forest communities of southern Wisconsin. *Ecol Monogr* **27**: 325–349.

Bryant D, Moulton V. (2004). Neighbor-net: an agglomerative method for the construction of phylogenetic networks. *Mol Biol Evol* **21**: 255–265.

Gilbert JA, Field D, Huang Y, Edwards R, Li W, Gilna P *et al.* (2008). Detection of large numbers of novel sequences in the metatranscriptomes of complex marine microbial communities. *PLoS One* **3**: e3042.

Gilbert JA, Field D, Swift P, Newbold L, Oliver A, Smyth T *et al.* (2009). The seasonal structure of microbial communities in the Western English Channel. *Environ Microbiol* **11**: 3132–3139.

Goodall DW. (1964). A probabilistic similarity index. *Nature* **203**: 1098.

Goodall DW. (1966). A new similarity index based on probability. *Biometrics* **22**: 882–907.

Huson DH, Auch AF, Qi J, Schuster SC. (2007). MEGAN analysis of metagenomic data. *Genome Res* **17**: 377–386.

Lebart L, Morineau A, Félon JP. (1979). *Traitement des Donndées Statistiques - Méthodes et Programmes*. Dunod: Paris.

Legendre P, Legendre L. (1998). *Numerical Ecology*. English edn. **20**: i–xv, 1-853.

Odum EP. (1950). Bird populations of the highlands (North Carolina) plateau in relation to plant succession and avian invasion. *Ecology* **31**: 587–605.

Rao CR. (1995). A review of canonical coordinates and an alternative to correspondence analysis using Hellinger distance. *Qüestiió (Quaderns d'Estadística i Investigació operativa)* **19**: 23–63.

Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S *et al.* (2007). The sorcerer II global ocean sampling expedition: Northwest Atlantic through Eastern Tropical Pacific. *PLoS Biol* **5**: 398–431.

Seshadri R, Kravitz SA, Smarr L, Gilna P, Frazier M. (2007). CAMERA: a community resource for metagenomics. *PLoS Biol* **5**: 394–397.

Sogin ML, Morrison HG, Huber JA, Mark Welch D, Huse SM, Neal PR *et al.* (2006). Microbial diversity in the deep sea and the underexplored 'rare biosphere'. *Proc Natl Acad Sci USA* **103**: 12115–12120.

Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA *et al.* (2004). Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**: 66–74.

Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V *et al.* (2008). Database resources of the national center for biotechnology information. *Nucleic Acids Res* **36**: D13–D21.

Supplementary Information accompanies the paper on The ISME Journal website (http://www.nature.com/ismej)