**Phylogenetics**

**RAMI: a tool for identification and characterization of phylogenetic clusters in microbial communities**

Thomas Pommier\(^1,\text{*} ,\) \(^\dagger\), Björn Canbäck\(^2,\text{*}\), Per Lundberg\(^3\), Åke Hagström\(^1\) and Anders Tunlid\(^4\)

\(^1\)Department of Natural Science, Kalmar University, Kalmar, \(^2\)Björn Canbäck Bioinformatics, Vindögatan 66, SE-257 33 Rydebäck, \(^3\)Department of Theoretical Ecology and \(^4\)Department of Microbial Ecology, Lund University, Lund, Sweden

Received on September 06, 2008; revised and accepted on January 22, 2009

Advance Access publication February 17, 2009

Associate Editor: Martin Bishop

**ABSTRACT**

**Motivation:** The most common approach to estimate microbial diversity is based on the analysis of DNA sequences of specific target genes including ribosomal genes. Commonly, the sequences are grouped into operational taxonomic units based on genetic distance (sequence similarity) instead of genetic change (patristic distance). This method may fail to adequately identify clusters of evolutionary related sequences and it provides no information on the phylogenetic structure of the community. An ease-of-use web application for this purpose has been missing.

**Results:** We have developed RAMI, which clusters related nodes in a phylogenetic tree based on the patristic distance. RAMI also produces indices of cluster properties and other indices used in population and community studies on-the-fly.

**Availability:** RAMI is licensed under GNU GPL and can be run or downloaded from http://www.acgt.se/RAMI.

**Contact:** tpommier@univ-montp2.fr; bcanback@acgt.se

**Supplementary information:** http://www.acgt.se/RAMI/SupplInfo

**1 INTRODUCTION**

DNA sequencing has become the major method for characterizing the diversity of microorganisms in nature. Recently, this approach has been reinforced by the introduction of novel techniques for ultra-high-throughput DNA sequencing (Sogin et al., 2006). The molecular techniques have revealed an immense genetic diversity of microorganisms, most of which is not yet characterized. Typically, the data consist of sequences from a given target gene including ribosomal genes. To be analyzed, the sequences are commonly clustered into operational taxonomic units (OTUs) using an arbitrary limit of sequence similarity. While such groupings have successfully been used for analyzing the structure of microbial communities in numerous studies, potentially valuable information concerning the relationships among sequences and the phylogenetic structure of the communities is lost (Bohannan and Hughes, 2003).

In this report, we present a new tool—RAMI (i.e. the Latin form of ‘branches’) that aims to identify and classify groups or ‘clusters’ in phylogenetic trees, based on the so-called patristic distance (i.e. the branch lengths) and to characterize their structure, variations and relationships. RAMI can be combined and integrated with a number of different software programs for analyzing the phylogenetic structure of ecological communities and populations (Fig. 1). When run as a web application, RAMI provides an ease-of-use tool to analyze datasets which eliminates the need of downloading, installing and running programs locally. We demonstrate the usefulness of RAMI using a dataset of 16S ribosomal RNA (rRNA) genes from communities of marine bacterioplankton. RAMI could be used for characterizing the cluster structures in trees constructed from any type of data and the tool could be applied for characterizing the phylogenetic patterns of diversity of all kinds of organisms.

**2 METHODS**

Available clustering algorithms use genetic distances between sequences to build clusters. The genetic distance is calculated from scores that may be produced in various ways. Patristic distances represent the amount of genetic changes between sequences. In a phylogenetic tree in the form of a phylogram patristic distances correspond to the lengths of the branches. Very few tree reconstruction programs output patristic distance matrices but nearly all have the option to save the tree file in Newick or related formats. RAMI uses such files as input file to calculate the patristic distances between both internal and external nodes. Using a single-linkage algorithm, RAMI then clusters sequences into OTUs that are found within a patristic distance set by the user. Once the clusters are defined, a number of indices are calculated (Fig. 2).

The first three indices derive from comparisons of nodes within sequence clusters: $X_{\text{distance}}$, the average patristic distance between external nodes; $X_{\text{depth, nearest}}$, the average patristic distance from external nodes to their adjacent ancestral nodes and $X_{\text{depth, deepest}}$, the average patristic distance from external nodes to the base node in the cluster (Fig. 2a). The last three indices derive from comparisons between sequence clusters: $Y_{\text{distance}}$, the average patristic distance between clusters; $Y_{\text{depth, nearest}}$, the patristic distance from the cluster to the adjacent ancestral node and $Y_{\text{depth, deepest}}$, the patristic distance from the cluster to the root node of the tree (Fig. 2b). Note that the names of the $X$ and $Y$ indices indicate that the measurements are similar, but at different scales. The value of the $Y_{\text{depth, deepest}}$ index depends on the choice of outgroup. More distant outgroups will produce higher values. The user has the option to exclude outgroups from the analysis.

The averages of the $Y_{\text{depth, nearest}}$ indices correlate to the mean nearest phylogenetic neighbor distance (MNND) calculated by the construct module of PHYLOCOM (Webb et al., 2008). The difference is that RAMI uses...
is important to get a proper value of the
RAMI has the option to remove outgroup sequences from the analysis, which
indices. The value of $X_{\text{depth, nearest}}$, $X_{\text{depth, deepest}}$ and $X_{\text{Distance}}$ indices. The value of $X_{\text{depth, deepest}}$ is dependent on the choice of outgroup.
RAMI has the option to remove outgroup sequences from the analysis, which is important to get a proper value of the $Y_{\text{distance}}$ index. Note that cluster 1 is paraphyletic.

the distance to the adjacent node while construct uses the distance to
the nearest OTU. The average of the $Y_{\text{distance}}$ indices will be very similar or identical to the mean phylogenetic distance (MPD) also calculated by

3.1 Comparison of cluster assemblies produced by
RAMI and analogous programs

To assess the quality of RAMI’s clustering approach, we compared assemblies of clusters produced by three different clustering
algorithms, RAMI, DOTUR (Schloss and Handelsman, 2005) and
BLASTclust using 269 full-length $\gamma$-proteobacterial 16S rRNA sequences from the manually curated Greengenes database core set
(DeSantis et al., 2006). Distance thresholds for respective algorithm
were set to produce the same number of clusters for at least
two of the methods starting from an assembly consisting of only
singletons (for specific settings see Supplementary Material). As
mentioned, the distance measure used in RAMI is the patristic
distance while DOTUR was originally designed to use a similarity
distance matrix in DOTUR. This work-around follows several steps:
(i) a phylogenetic tree must be calculated with PAUP (Swofford,
2003), RaxML (Stamatakis, 2005). However, it is also possible to input a patristic
distance matrix in DOTUR. This work-around follows several steps:
(i) a phylogenetic tree must be calculated with PAUP (Swofford,
2003), RaxML (Stamatakis et al., 2005) or similar software; (ii)
patristic distances must be calculated from the tree file using, e.g.
PATRISTIC (Fourment and Gibbs, 2006), the phydist module of

PHYLLOCOM. However, to allow equivalent MNND and MPD calculations in
PHYLLOCOM, a single representative sequence for each sequence cluster
has to be determined and used.

RAMI outputs four files: (i) a list of all sequences and the clusters (OTUs)
they belong to; (ii) a list of clusters, their indices and sequence abundances;
(iii) a file with distances between connecting nodes, similar to the output from
PAUP (Swofford, 2003); and (iv) a two-column matrix file with distances
between external nodes, similar to the output from PATRISTIC (Fourment
and Gibbs, 2006) or the phydist module of PHYLLOCOM (Webb et al.,
2008).

When run as a web-application, RAMI first outputs the files described
above, and then the user can access a number of analyses and visualizations
tools: (i) a randomized Chaos richness estimator (Lee and Chao, 1994)
curve of the OTUs; (ii) a randomized accumulation curve of the OTUs;
(iii) the Shannon index and evenness value (Shannon and Weaver, 1949);
(iv) the visualization of the tree including the OTUs by automatic
submission to the online tool iTOL (Letunic and Bork, 2007); (v) a
two-column matrix file with distances between all nodes, both internal
and external, which is unique to RAMI; (vi) the creation of a new tree
with OTUs as external nodes, which facilitates the visualization of trees
with large number of external nodes. This new tree may be based on
a matrix containing the distances between the base nodes of all OTUs
marked with an open circle in Fig. 2) or these distances added with the
average value of the $X_{\text{depth, deepest}}$ index. Again, the user has the option
to visualize the tree in iTOL together with circles with areas representing
abundances of sequences included in each OTU; (vii) a compressed file with
clusters and their sequences; (viii) a FASTA file including the consensus
sequences for each cluster as calculated by cons from the EMBOSS
package (Rice et al., 2000); and finally (ix) the calculation of the Net
Relatedness Index (NRI) and the Nearest Taxa Index (NTI) developed
for the construct module in the PHYLOCOM package (Webb et al.,
2008). These indices measure the degree of phylogenetic clustering or
overdispersion. It should be reminded that RAMI calculates these indices
based on sequence clusters that is here treated as OTUs. To our knowledge,
no other software is able to calculate these indices for sequence clusters,
which can be a major advantage when analyzing samples with many closely
related sequences. We demonstrate such application of RAMI to marine
bacterioplankton communities assessed by 16S rRNA gene sequences in
Section 3.2.
Some of these are presented in the lower panel. 

Fig. 3. Comparison of different clustering algorithms based on 269 full length \(\gamma\)-proteobacterial rRNA sequences from Greengenes manually curated core set (DeSantis et al., 2006). Identical cluster assemblies have a value of 0. In all pair-wise comparisons, distance thresholds for the respective algorithms have been set to produce the same number of clusters. Some of these are presented in the lower panel.

PHYLOCOM (Webb et al., 2008) or RAMI; and finally, the output must be reformatted to the phylip format (Fig. 1). This work-around is also included in the analysis.

BLASTclust on the other hand, uses the similarity measure identity for clustering and we have therefore defined the distance as the value of 1—identity (Fig. 3). Differences in cluster assemblies were estimated with the variation of information (VI) metric (Meilä, 2003, 2007). This metric compares two sets of clusters assembled from the same input (in this case sequences) by different algorithms or settings. Though we based our analysis on similar number of clusters, the VI index does not require that the same number of clusters is produced from the two datasets. Identical assemblies will have a VI value of 0 and the higher the dissimilarity the higher the VI value.

RAMI and DOTUR produced identical clusters (i.e. VI = 0) when distance thresholds were very low, i.e. when DOTUR distance was less than 0.03 (Fig. 3). In contrast, clusters assemblies produced with BLASTclust differed from the two other methods already at very low distance thresholds. Increasing distance thresholds resulted in higher differences between the assemblies, at least up to a cluster size of 152. To produce the same number of clusters (but with different information content), DOTUR and BLASTclust used very close distance thresholds, while RAMI needed to be run with approximately twice these thresholds. These results imply that when analyzing short and conserved sequences, RAMI and DOTUR should produce very similar cluster assemblies. However, when clustering longer or less conserved sequences, differences between the two algorithms should also be observed at low distance thresholds. This is especially true if the model of sequence evolution is not properly approximated by the parameters used in dnadist that creates the underlying matrix used in DOTUR. Indeed, dnadist relies on a user-supplied coefficient of substitution rate of variation when applying a gamma distribution and does not allow for a general time-reversible model with six substitution rates. Additionally, if the model of sequence evolution involves asymmetric substitution rates and heterogeneous G + C contents, incorrect clustering may occur when using similarity-based assemblies (Supplementary Material, Fig. S1).

When supplying DOTUR with patristic distances calculated by PATRISTIC from the same input tree as used in RAMI, the two software programs produced identical results. This was also true for a set of 1012 bacterial sequences of the single copy recA gene, which validates the single-linkage algorithms used in the two programs (data not shown).

3.2 Application to bacterial ribosomal DNA sequences from the marine environment

A usual observation when analyzing genetic markers from environment samples is the occurrence of numerous closely related sequences, which are often referred to as microdiverse sequence clusters. Microdiversity within ribosomal RNA (rRNA) genes has been reported in several microorganisms in the marine environment. In an original approach to explain microdiversity patterns, Acinas et al. (2004) examined the occurrence of microdiverse clusters in bacterial communities from one coastal environment sample located in the Plum Island Sound. They found a large number of closely related phyotypes (\(\geq 99\%\) similar) that were independently but variably distributed among taxonomic lineages.

To complete and compare this study with recent data, we added data from seven clone libraries from Pommier et al. (2007). The samples were collected from different localities (Sargasso Sea and offshore Cape Town, Concepción de Chile, Fiji, Hawaii, San Diego and Sydney) spread around the world.

A strict selection for accurate sequences nominated 2878 sequences from the seven locations from Pommier et al., and 1081 sequences from Acinas et al. (see Supplementary Material for a description of the method). All sequences were aligned using the online tool from Greengenes (DeSantis et al., 2006). The total alignment was divided into two datasets, one with alignments for each location and one with alignments for each major taxonomic group. From these alignments, we used the maximum likelihood method as implemented in RAxML (Stamatakis et al., 2005) to build phylogenetic trees. (Please consult Supplementary Material for specific settings of various programs.)

In RAMI, a microdiverse cluster will be defined as a group of nodes that is separated from other nodes with a patristic distance less than a given cutoff value. Obviously, the level of threshold value will determine the number of microdiverse clusters identified within the analyzed community. Using a patristic distance cutoff value of 0.01 substitutions per nucleotide, RAMI could outline from 92 to 261 microdiverse clusters, with an average of 174 clusters for each community. An increase of patristic distance to 0.03 or 0.05 dropped the average number to 128 and 106, respectively. Using the clusters defined by RAMI, we recovered the same features of the structure of marine bacterioplankton communities as when we defined OTUs with a score based cutoff (Pommier et al., 2007). For example, the fraction of all microdiverse clusters within a locality...
that was endemic remained constant (86%) across all localities and irrespective of community size (taxon richness).

To assess the level of phylogenetic clustering and overdispersion, respectively, in the different communities, RAMI has the option to calculate the NRI and NTI indices based on sequence clusters (OTUs), which is not possible to do in PHYLOCOM (Webb et al., 2008). NRI measures the degree of clustering in a phylogenetic tree when comparing a community with the regional pool (the entire tree), while NTI measures the degree of clustering of external nodes. We added the dataset from Plum Island Sound (Acinas et al., 2004) to our analysis to address the issue of phylogenetic structuring in marine samples collected worldwide. Three of the communities, Sargasso Sea, Plum Island Sound and Fiji, showed a high degree of phylogenetic clustering (significant at the \( P = 0.05 \) level). In contrast, the Sydney community showed a high degree of overdispersion (significant at the \( P = 0.10 \) level) (Fig. 4).

To illustrate the geographic distributions and sizes of the microdiverse clusters produced by RAMI in the major taxonomic groups of the data from Pommier et al., we created plots of microdiverse clusters colored according to their sampling sites (Fig. 5). When comparing the plot of the phylum *Cyanobacteria* (Fig. 5a) with the one of *Cyanobacteria* (Fig. 5b), it was evident that the two phyla were differently structured both from a geographic and a phylogenetic perspective. While the large majority of cyanobacterial sequences were found in three major clusters of *Prochlorococcus* and *Synechococcus*, sequences from *Verrucomicrobiae* tended to be more evenly distributed. Sequences from *Verrucomicrobiae* produced clusters of sequences based on genetic change (the so called patristic distance) instead of a score-based genetic distance.

One interpretation of such tree topology may be that the organisms carrying these sequences have escaped the effects of selective sweeps of their ancestral populations (Cohan, 2001).

Table 1 presents the six indices produced by RAMI while building clusters of sequences belonging to the same taxonomic group. As expected from the graphical views (Fig. 5), the average cluster size was larger for *Cyanobacteria* (5.3 sequences) than for *Verrucomicrobiae* (2.1). On average, the \( Y_{\text{depth.nearest}} \) indices were 0.045 for *Verrucomicrobiae* and 0.037 for *Cyanobacteria*. This was also in agreement with the visual impression (see above and Fig. 5) that the *Verrucomicrobiae* tree had a number of long internal branches. On average, the \( Y_{\text{distance}} \) indices were 0.49 for *Verrucomicrobiae* and 0.22 for *Cyanobacteria*. Considering the larger number of clusters in *Cyanobacteria*, these are surprising values. We conclude that the cyanobacterial clusters were less divergent to each other than clusters from *Verrucomicrobiae*. The average of the \( Y_{\text{depth.deepest}} \) indices were 0.34 for *Verrucomicrobiae* and 0.13 for *Cyanobacteria*. Again, these values correspond to the visual impression of the two trees: in the cyanobacterial tree, the OTUs were in general very close to the base of the tree. It should be emphasized that in trees where no outgroup has been assigned, like the ones in this study, the index is strictly dependent on which root is used for tree visualization. If the two phyla had been assigned to the same outgroup taxa, a comparison of the index values would indicate the amount of sequence evolution for respective phylum since their divergence. The reason for not including outgroups in this study is that highly variable sequence positions may be masked out by including distantly related outgroups like *Archaea*. This is especially true when using relatively short sequences as in this case.

The ‘\( X \)' indices measure properties within clusters and will always be 0 in singleton clusters. They are thus best suited for comparisons between clusters with similar sizes since they are dependent on sequence abundances. The two largest clusters (the top left and the bottom right clusters in Fig. 5b) in *Cyanobacteria* are well suited for this type of analysis. Sequence abundances for the two clusters are 71 and 68, respectively. The \( X_{\text{depth.deepest}} \) index value for the larger cluster was 0.0097 but only 0.0024 for the smaller one. This may indicate a more recent divergence of sequences in the smaller cluster. The corresponding figure for the \( X_{\text{depth.nearest}} \) index was 0.00042 for both clusters. This shows that sequences diverged at the same rate in both clusters. Taken together with the values of the \( X_{\text{depth.deepest}} \) index, it can be concluded that evolution of the smaller cluster is more like a quick radiation while sequences in the larger cluster have evolved in small progressive steps. The values of the \( X_{\text{distance}} \) indices were 0.0052 for the larger cluster and 0.0027 for the smaller, indicating that sequences in the smaller cluster were less divergent than in the larger cluster.

4 DISCUSSION

We have developed a software tool called RAMI to identify and characterize clusters derived from phylogenetic trees (i.e. phylograms). RAMI’s main application will probably be to create and characterize clusters based on phylogenies constructed from sequence data, but it could be used for any data that is meaningful to display in a phylogenetic tree. RAMI accepts various types of input tree files, produced by any phylogenetic method. RAMI produces clusters of sequences based on genetic change (the so called patristic distance) instead of a score-based genetic distance,
which should result in accurate and evolutionary robust clustering. We argue that the measure of patristic distance used in RAMI is more correct from a theoretical standpoint than score based distances used in other algorithms, in analogy with the fact that the likelihood methods are often preferred to the distance methods in tree reconstruction. The two approaches will give similar results when analyzing sequences with low rates of evolution and similar base compositions.

As pointed out above, it is a possible to use DOTUR with patristic distances if these are calculated from a tree with tools such as phydist from PHYLOCOM, PATRISTIC (which should not be used together with larger datasets due to memory limitations) or RAMI. The output has to be reformatted to a matrix in the phylip-format. This workaround requires some programming skills and additional use of software. Unlike DOTUR, RAMI computes a number of indices that describe cluster properties. Apart from this, RAMI also creates randomized OTU accumulation curves and randomized Chao1 estimator curves (Lee and Chao, 1994), computes the Shannon index (Shannon and Weaver, 1949), visualizes clusters in phylogenetic trees with the aid of iTOL (Letunic and Bork, 2007), calculates the NRI and NTI indices and produces aggregated trees with clusters as OTUs.

Web applications like RAMI have the advantage of requiring insignificant computer resources from the client, that no installation is required, and that user always has access to the latest version of the program. Not least important is that utilizing the software becomes platform independent.

We foresee that RAMI with its ease-of-use interface will be a valuable tool for researchers who want to analyze phylogenetic structure of microbial communities. Phylogenies of DNA sequences generated from environmental samples of microbial communities typically contain hierarchies of clusters and sub-clusters within clusters, and the relationships between sequence clusters, bacterial species and ecotypes (i.e. ecologically distinct populations) have been intensively discussed (Cohan, 2001; Gevers et al., 2005). While present methods use universal thresholds to identify OTUs, RAMI recognizes clearly resolved clusters of sequences based on genetic change in phylogenetic trees. The delineation of clusters and sub-clusters and accordingly their sizes, will depend on the threshold settings. To confirm whether the recognized clusters represent
distinct species and/or ecotypes additional analyses are required. For example, a single gene might have too few variable nucleotide sites to resolve very similar species or ecotypes. Information from several genes might also be required to identify cases of recombination that may distort the assignments of species to clusters of single gene sequences. Ecological approaches are needed to identify ecotypes among sequence clusters. In such cases, clusters obtained in RAMI could provide a guide for the selection of isolates for ecological studies.

5 IMPLEMENTATION

RAMI is written in PERL. The standalone version should work on any operating system running PERL but was developed and tested with Linux as operating system. Code for standalone usage can be downloaded from the web site (http://www.acgt.se/online.html).

Table 1. RAMI indices for describing the microdiverse clusters identified in marine Verrucomicrobiae and Cyanobacteria

| Cluster | Abundance | $X_{\text{distance}}$ | $X_{\text{depth.nearest}}$ | $X_{\text{depth.deepest}}$ | $Y_{\text{distance}}$ | $Y_{\text{depth.nearest}}$ | $Y_{\text{depth.deepest}}$ |
|---------|-----------|---------------------|--------------------------|--------------------------|---------------------|--------------------------|--------------------------|
| Verrucomicrobiae |
| 1 | 5 | 0.008727 | 0.002399 | 0.004804 | 0.480865 | 0.011001 | 0.264861 |
| 2 | 10 | 0.004843 | 0.000000 | 0.006593 | 0.457065 | 0.081461 | 0.111974 |
| 3 | 4 | 0.005598 | 0.001576 | 0.004198 | 0.457065 | 0.081461 | 0.111974 |
| 4 | 4 | 0.002858 | 0.000000 | 0.002143 | 0.533413 | 0.002793 | 0.43239 |
| 5 | 2 | 0.002344 | 0.001172 | 0.001172 | 0.468564 | 0.002793 | 0.43239 |
| 6 | 3 | 0.000003 | 0.000001 | 0.000002 | 0.473533 | 0.051559 | 0.199619 |
| 7 | 2 | 0.000002 | 0.000001 | 0.000002 | 0.467676 | 0.062627 | 0.432225 |
| 8 | 5 | 0.001952 | 0.000975 | 0.000977 | 0.446373 | 0.000977 | 0.432225 |
| 9 | 2 | 0.003489 | 0.001745 | 0.001745 | 0.434259 | 0.002770 | 0.395029 |
| 10–26 | 1 | 0 | 0 | 0 | 0.486456 | 0.044597 | 0.338472 |

Cyanobacteria

| Cluster | Abundance | $X_{\text{distance}}$ | $X_{\text{depth.nearest}}$ | $X_{\text{depth.deepest}}$ | $Y_{\text{distance}}$ | $Y_{\text{depth.nearest}}$ | $Y_{\text{depth.deepest}}$ |
|---------|-----------|---------------------|--------------------------|--------------------------|---------------------|--------------------------|--------------------------|
| 1 | 71 | 0.005218 | 0.000418 | 0.009660 | 0.135964 | 0.000001 | 0.19574 |
| 2 | 43 | 0.004877 | 0.001027 | 0.005124 | 0.155956 | 0.029753 | 0.038109 |
| 3 | 68 | 0.002724 | 0.000417 | 0.002380 | 0.130008 | 0.000000 | 0.000000 |
| 4 | 11 | 0.005435 | 0.000258 | 0.005030 | 0.139756 | 0.011151 | 0.019505 |
| 5 | 3 | 0.004142 | 0.002070 | 0.002072 | 0.153375 | 0.012266 | 0.035362 |
| 6 | 3 | 0.000003 | 0.000001 | 0.000002 | 0.156087 | 0.011149 | 0.036184 |
| 7 | 4 | 0.003405 | 0.000811 | 0.002149 | 0.283763 | 0.023241 | 0.262545 |
| 8 | 9 | 0.000005 | 0.000001 | 0.000005 | 0.420673 | 0.075142 | 0.410896 |
| 9 | 4 | 0.004759 | 0.001431 | 0.002854 | 0.156194 | 0.010309 | 0.037752 |
| 10 | 2 | 0.009941 | 0.004970 | 0.004970 | 0.358074 | 0.093552 | 0.205092 |
| 11 | 4 | 0.000707 | 0.002702 | 0.002702 | 0.286077 | 0.022456 | 0.266408 |
| 11–49 | 1 | 0 | 0 | 0 | 0.220065 | 0.044597 | 0.127355 |

5.3 0.004325b 0.001282b 0.003473b 0.220065 0.044597 0.127355

acknowledgements

We would like to thank Morten Krogh, BMC, Lund University, for help with time optimization of the RAMI code and implementation of the VI metric, and Josephine Leflaive, Université de Montpellier II, for her suggestions in statistical analyses.

funding: Swedish Research Council (VR); Swedish Research Council for Environment, Agricultural Science and Spatial Planning (FORMAS).

Conflict of Interest: none declared.

REFERENCES

Acinas,S.G. et al. (2004) Fine-scale phylogenetic architecture of a complex bacterial community. Nature, 430, 551–554.

Bohannan,B.J. and Hughes,J. (2003) New approaches to analyzing microbial biodiversity data. Curr. Opin. Microbiol., 6, 282–287.

Cohan,F.M. (2001) Bacterial species and speciation. Syst. Biol., 50, 513–524.

DeSantis,T.Z., et al. (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl. Environ. Microbiol., 72, 5069–5072.

Felsenstein,J. 2005. PHYLIP (Phylogeny Inference Package) version 3.6. Department of Genome Sciences, University of Washington, Seattle.

Fourment,M. and Gibbs,M. (2006) PATRISTIC: a program for calculating patristic distances and graphically comparing the components of genetic change. BMC Evol. Biol., 6, 1.

Govers,D et al. (2005) Opinion: re-evaluating prokaryotic species. Nat. Rev. Microbiol., 3, 733–739.
Lee, S.-M. and Chao, A. (1994) Estimating population size via sample coverage for closed capture-recapture models. *Biometrics*, **50**, 88–97.

Letunic, I. and Bork, P. (2007) Interactive Tree Of Life (iTOl): an online tool for phylogenetic tree display and annotation. *Bioinformatics*, **23**, 127–128.

Meilã, M. (2003) Comparing clusterings by the variation of information. In *Proceedings of the Sixteenth Annual Conference of Computational Learning Theory (COLT)*. Springer, pp. 173–187.

Meilã, M. (2007) Comparing clusterings - an information based distance. *J. Multivar. Anal.*, **98**, 873–895.

Pommier, T. *et al.* (2007) Global patterns of diversity and community structure in marine bacterioplankton. *Mol. Ecol.*, **16**, 867–880.

Rice, P. *et al.* (2000) EMBOSS: the European molecular biology open software suite. *Trends Genet.*, **16**, 270–277.

Schloss, P.D. and Handelsman, J. (2005) Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. *Appl. Environ. Microbiol.*, **71**, 1501–1506.

Shannon, C.E. and Weaver, W. (1949) *The Mathematical Theory of Communication*. University of Illinois Press, Urbana.

Sogin, M.L. *et al.* (2006) Microbial diversity in the deep sea and the underexplored “rare biosphere”. *Proc. Natl Acad. Sci. USA*, **103**, 12115–12120.

Stamatakis, A. *et al.* (2005) RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. *Bioinformatics*, **21**, 456–463.

Swofford, D.L. (2003) *PAUP*+: *Phylogenetic Analysis Using Parsimony (*And Other Methods)*. Version 4. Sinauer Associates, Sunderland, MA.

Webb, C.O. *et al.* (2008) *Phylocom: software for the analysis of phylogenetic community structure and trait evolution*. *Bioinformatics*, **24**, 2098–2100.