New high-throughput biotechnologies for sampling the microbial ecological diversity of the oceans: the informatics challenge

Carmen Palacios[1,2], Bertil Olsson[3], Philippe Lebaron[2], Mitchell L. Sogin[3]

[1]Max Planck Institute for Marine Microbiology, Celsiusstr. 1
28359 Bremen, German

[2]Observatoire Océanologique Banyuls, Laboratoire de Microbiologie
66651 Banyuls-sur-mer, France
E-mail: carmen.palacios@obs-banyuls.fr

[3]Marine Biological Laboratory. 7 MBL St.
Woods Hole, 02543 MA, USA

Abstract

Microorganisms account for most of Earth's biodiversity. They mediate key biogeochemical processes and serve pivotal functional roles in complex ecosystems, yet little is known about mechanisms responsible for the formation of microbial ecological diversity patterns. High-throughput molecular biology provides a powerful tool for measuring and monitoring patterns of microbial diversity. SARST-V6 (Serial Analysis of V6 Ribosomal Sequence Tags) is a promising technology that uses short DNA sequence tags to fingerprint the composition of microbial communities. To efficiently interpret the large amount of diversity information generated by this high-throughput technique we have made significant improvements to our SARST-V6 data acquisition and analysis informatics tools, which is now available through the WEB portal http://www.obs-banyuls.fr/UMR7621/SARST-V6.

Keywords: High-throughput microbial community analysis; Microbial ecological diversity; SARST-V6; Sunken woods.

Background information

Understanding patterns of variation in microbial populations is of great importance because these relatively simple organisms account for the majority of biodiversity on earth where they mediate key processes that sustain all forms of life. For instance, microorganisms may represent as much as 90% of biomass in marine systems where they serve key roles in remineralization of carbon, with and without oxygen, nitrogen cycling, and biogeochemical transformations of sulfur, iron and manganese (Kirchman, 2000). Microbial ecological diversity investigations in concert with detailed descriptions of environmental parameters promise to unveil new insights about interactions between microorganisms and their habitats (Green *et al.*, 2004; Horner-Devine *et al.*, 2004). Early studies of microbial molecular diversity relied upon sequence analyses of ribosomal RNAs (rRNAs) or their coding regions (Hugentzholtz *et al.*, 1998). Although rich in

information, these investigations are expensive to perform and usually provide no reliable abundance data of the different kinds of organisms at a study site because of its limited throughput. To address this problem, investigators turned their attention to relatively rapid profiling methods such as terminal restriction fragment length polymorphisms (T-RFLP, Moeseneder *et al.*, 1999) or denaturing gradient gel electrophoresis (DGGE, Muyzer *et al.,* 1993). These techniques provide estimates of the relative number of specific rRNA amplicons generated by polymerase chain reaction (PCR) experiments but without accompanying DNA sequence analyses, they lack the ability to identify specific phylotypes in a given community. Moreover, since these methods simply measure relative amounts of nucleic acid from different organisms in a sample, fluorescence in-situ hybridization (FISH, Amann *et al.,* 1995) remains the most reliable method to determine the number of probed organisms. However, FISH is not a high-throughput method and it only detects organisms with rRNAs that hybridize with the designed probes.

New methods with the potential to overcome these technical difficulties are at various stages of development (Bertilsson *et al.*, 2002; Neufeld *et al.*, 2004; Kysela *et al.*, 2005). They exploit the intrinsic phylogenetic information contained in relatively short (30-150 base pairs), genetically hypervariable regions of the ribosomal RNA molecule to extract phylotype information directly from sequencing. The advantage of these technologies for ecological diversity studies is that they allow detection of all organisms present in natural samples through high-throughput sequencing while at the same time providing estimates of their relative numbers. Thus they avoid the difficulties and assumptions associated with estimating relative abundances of organisms based upon integration of band intensities generated by fingerprinting methods like DGGE or TRFLP. The high throughput generation of short sequence tags for studies of microbial diversity requires the capability to process large amounts of sequence data. In this communication we outline improvements to our informatics treatment of data from Serial Analysis of V6 Ribosomal Sequence Tags methodology (SARST-V6, Kysela *et al.*, 2005).

The informatics challenge: SARST-V6 pipeline

SARST-V6 is a molecular method that draws upon information-rich DNA sequence analysis of the 16S rRNA, while providing higher throughput and efficiency than standard small subunit ribosomal DNA sequencing protocols. The technique is modelled after serial analysis of gene expression (SAGE), which describes relative expression levels for genomic tags in mRNA populations (Velculescu *et al.*, 1995). SARST-V6 produces sequences of large concatemers of PCR-amplified ribosomal sequence tags (RSTs) from homologous V6 hypervariable regions (Kysela *et al.*, 2005). This strategy increases by at least 6-fold the yield of information about different PCR amplicons in a single sequence relative to the traditional sequencing of a single rRNA amplicon in each reaction. To extract biodiversity information from the concatemer sequences, it is necessary to identify the boundaries of each RST. Comparison against a comprehensive rRNA gene database identifies the taxonomic assignment of individual RSTs.

The flow chart in Figure 1 outlines the SARST-V6 pipeline. A pipeline consists of several scripts and programs that carry out a series of bioinformatics steps required to

process data. Our pipeline aims to extract ecological diversity information from SARST-V6 data analysis. In the flow chart, customized scripts to process SARST-V6 concatemer sequences into individual RSTs are intermingled with available software (usually freeware) to analyze sequence data. All scripts particular to SARST-V6 contributed by this communication are available upon request.

From chromatogram files to sequence FASTA files

The first step in the SARST-V6 pipeline (Fig. 1) is to convert chromatograms into PHD format sequencing files using PHRED (freeware available at http://www.phrap.com/background.htm). PHD files contain not only the base pair sequence information but also the quality of each called base. To automatically trim PHD files from vector sequence and low quality reads, we use LUCY (freeware at http://www.tigr.org) with default parameters with the exception that minimum sequence length is set to 20 in order to capture single tag sequences. PHD files are converted into regular FASTA format using PHD2FASTA (http://www.phrap.com/background.htm).

From concatemer sequences to ribosomal sequence tags (RSTs)

The second stage of the pipeline identifies the boundaries of individual tags and parses the concatemer into RSTs (Fig. 1). This script recognizes imperfect punctuations that arise because of sequencing errors or failure of the type II restriction enzymes to accurately cut at their predicted cleavage sites. In addition to imperfect punctuations in SARST concatemers, there can be other artifacts generated during DNA ligation or recombinant cloning. As part of this process the software generates a SARST file (Fig. 1) that contains all RSTs and marks those that reside at the beginning or end of the concatemer as well as artefactual and truncated tags. The SARST file provides a basis for making quality control decisions about the identity and integrity of RSTs (see below). The script to generate SARST files can be run interactively over the Internet for a single concatemer sequence (see Fig. 1 for output details. URL: http://www.obs-banyuls.fr/UMR7621/SARST-V6). However, to process larger amounts of data, it will be more efficient to download the programs and scripts for use on local LINUX computers, which are publicly available in the above URL.

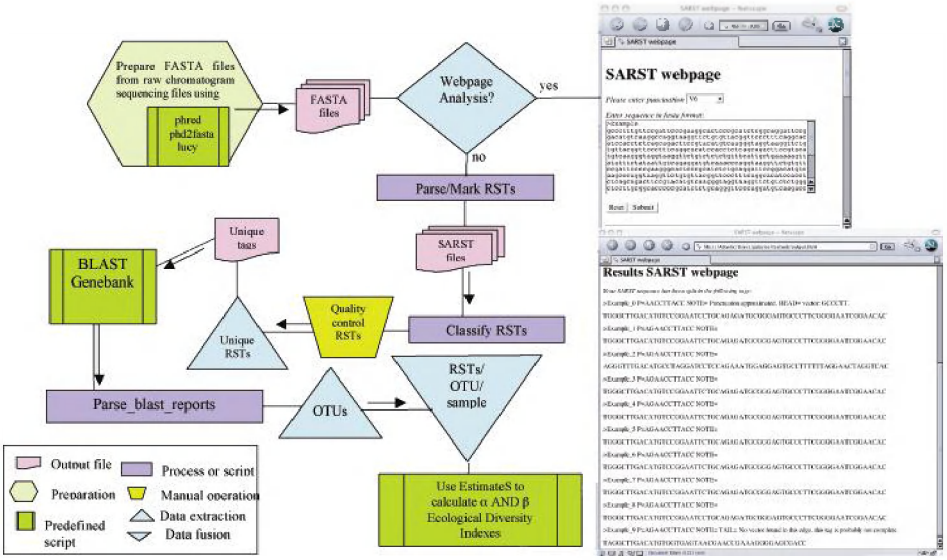

Fig. 1. SARST-V6 pipeline. This pipeline outlines the different stages of SARST-V6 sequence analysis (see text for details).

Forming integral RSTs free of artifacts

Quality control of parsed RSTs by manual inspection is accelerated through classification of RSTs according to the mark imprinted in the SARST file in previous step of the pipeline (Fig. 1). Thus we use a customized script to classify RSTs into different groups if their position in the concatemer was first, middle or last, whether or not vector boundaries are present in first and last tags and they are short, and/or if they have a particular artifact. Most of the RSTs will have no artifacts requiring no further processing. However, some RSTs in first or last position will be too short and therefore not complete or the whole SARST file lack perfectly punctuated tags. These groups of RSTs are eliminated from the dataset. RSTs that do not fall into these categories are inspected by eye for quality control. First, each classified group of RSTs according to a particular artifact is assembled into smaller, high similarity subgroups. Two programs that can assemble sequences are PHRAP (http://www.phrap.com/background.htm) and AlignIR (Technology University LI-COR, Inc). We use AlignIR 2.0.48 assemble algorithm with default parameters (minimum identity 70% and maximum successive failures 50) for this purpose. Without the assembly process, generating an alignment is not possible because of the genetic hypervariability of the region we are dealing with. Once a subgroup of RSTs is aligned, it is relatively easy to manually identify and remove the as well aligned particular artifact from all RSTs at a time.

Taxonomic affiliation of tags

After quality control of RSTs, the next step of the pipeline aims to determine their taxonomic affinity. We first pool all tags that are identical in their sequence into a unique RST (Fig. 1). A customized script will extract these unique RSTs while keeping track of how many of those tags occur in a particular environmental sample, what is necessary for estimating relative numbers of different sequence tags in the sample (see later).

The resulting unique RSTs are matched against publicly available sequence databases. We use BLAST program against nucleotide GeneBank database (http://www.ncbi.nlm.nih.gov). Resulting BLAST reports will return the organisms present in the database with the highest sequence affinity to each unique RST. We then parse these reports in a table in which unique RSTs are linked with the name of each most similar organism/s, BLAST score, e-value and sequence similarity to this organism's sequence.

Extracting OTUs and ecological analysis

Depending on the taxonomic resolution of interest (phylotype, genus, species, etc) a sequence similarity cut-off is chosen to group tags into Operational Taxonomic Units (OTUs). All tags matching a particular taxonomic group within that similarity cut-off are pooled together within an OTU.

By joining each of those extracted OTUs with the number of tags per OTU and per sample (this last value was registered in previous stage of the pipeline when extracting unique RSTs) we can then estimate species richness and evenness for each sample and β indexes of ecological diversity between samples using the freeware EstimateS (http://viceroy.eeb.uconn.edu/EstimateS).

SARST-V6 for ocean biodiversity studies: Future prospects

We have applied the SARST-V6 pipeline described in this communication to revisit the microbial diversity component of the water column and its correlation with physico-chemical parameters of the extremely acidic, high-metal laden Tinto River (Palacios *et al.*, unpublished data). Now that we have successfully explored the microbial diversity of relatively well-known environments using SARST-V6 (Kysela *et al.*, 2005; Palacios *et al.*, unpublished data), we can use this same methodology to characterize unexplored microbial communities like those that dwell sunken woods in deep waters. Sunken woods are very interesting deep-sea habitats from an evolutionary point of view as they might act as stepping-stones for chemosynthetic communities that inhabit hydrothermal vents and cold seeps (Smith *et al.*, 1989; Distel *et al.*, 2000). Our future application of SARST-V6 to sunken woods aims to explore the microbial patterns in these particular habitats and biogeochemical processes underlining them. It is likely that our results will give clues on the evolution of ocean biodiversity.

Conclusions

We have presented here our latest advances in data acquisition and analysis of SARST-V6 to illustrate the importance of Informatics when dealing with large datasets produced by high-throughput microbial community profiling methods. The SARST-V6 pipeline outlined in this communication will largely facilitate the analysis of the biodiversity generated using this sequencing technique. Computerization renders properly documented, well-organized datasets. The Ocean Biodiversity Informatics (OBI) conference statement summarizes that these characteristics allow data to be easily screened for errors, improving quality of released data. This has been our experience with SARST-V6 data analysis. Thanks to the Informatics' effort we are now ready to make publicly available our scripts and programs, hoping they will facilitate future studies of ocean biodiversity.

Acknowledgements

CP was supported by a postdoctoral stipend of the Max Planck Institute for Marine Microbiology (Bremen, Germany). We are grateful to Dave Kysela and Laura Shulman for help with the SARST-V6 script, and to Antje Boetius and Linda Amaral-Zettler for their support and encouragement.

References

Amann R.I., W. Ludwig and K.H. Schleifer. 1995. Phylogenetic identification and in situ detection of individual microbial cells without cultivation. Microbiololgy Reviews 59:143-169.

Bertilsson S., C.M. Cavanaugh and M.F. Polz. 2002. Sequencing-independent method to generate oligonucleotide probes targeting a variable region in bacterial 16S rRNAby PCR with detachable primers. Applied and Environmental Microbiology 68:6077-6086.

Distel D.L., A.R. Baco, E. Chuang, W. Morrill, C. Cavanaugh and C.R. Smith. 2000. Do mussels take wooden steps to deep-sea vents? Nature 403:725-726.

Green J.L., A.J. Homes, M. Westoby, I. Oliver, D. Briscoe, M. Dangerfield, M. Gillings and A.J. Beattie. 2004. Spatial scaling of microbial eukaryote diversity. Nature 432:747-750.

Horner-Devine M.C., M. Lage, J.B. Hughes and J.M. Bohannan. 2004. A taxa-area relationship for bacteria. Nature 432:750-753.

Hugenholtz P., B.M. Goebel and N.R. Pace. 1998. Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. Journal of Bacteriology 180:4765-4774.

Kirchman D.L. 2000. Microbial ecology of the oceans. John Wiley & Sons, New York. 542p.

Kysela D.T., C. Palacios and M.L. Sogin. 2005. Serial analysis of V6 ribosomal sequence tags (SARST-V6): a method for efficient, high-throughput analysis of microbial community composition. Environmental Microbiology 7:356-364.

Moeseneder, M.M., J.M. Arrieta, G. Muyzer, C. Winter and G.J. Herndl. 1999. Optimization of terminal-restriction fragment length polymorphism analysis for complex marine bacterioplankton communities and comparison with denaturing gradient gel electrophoresis. Applied and Environmental Microbiology 65:3518-3525.

Muyzer, G., E.C. de Waal and A.G. Uitterlinden. 1993. Profiling of complex microbial populations by denaturing gradient gel electrophoresis analysis of polymerase chain reaction-amplified genes coding for 16S rRNA. Applied and Environmental Microbiology 59:695-700.

Neufeld J.D., Y. Zhongtang, W. Lam and W.W. Mohn. 2004. Serial analysis of ribosomal sequence tags (SARST): a new high-throughput method for profiling complex microbial communities. Environmental Microbiology 6: 131-144.

Smith, C.R., H. Kukert, R.A Wheatcroft, P.A. Jumars and J.W. Deming. 1989. Vent fauna on whale remains. Nature 341:27-28.

Velculescu V.E., L. Zhang, B. Vogelstein and K.W. Kinzler. 1995 Serial analysis of gene expression. Science 270:484-87.