

MOLECULAR TOOLS FOR THE STUDY OF MARINE MICROBIAL DIVERSITY

L.K. Medlin, R. Groben, K. Valentin, Department of Biological Oceanography, Alfred Wegener Institute for Polar and Marine Research, Germany

Keywords:

AFLP, algae, biodiversity, clone library, DGGE, fingerprinting, FISH, microsatellites, molecular marker, molecular probe, oligonucleotide probes, PCR, phylogeny, phytoplankton, RAPD, RFLP, rRNA, SSCP, TGGE

Contents:

1. The Importance of Biodiversity Research in the Marine Environment
 2. What Questions Can be Answered Using Molecular Biology Techniques?
 3. Evaluating Marine Biodiversity by Sequence Analysis and Fingerprinting Methods
 - 3.1. Sequence Analysis
 - 3.1.1. Which Genes to Select?
 - 3.1.2. How to Generate Sequence Data?
 - 3.1.3. Determining Biodiversity in an Environmental Sample by Sequence Analysis
 - 3.1.4. Analysing Sequences for Determining Phylogenies and Biodiversity
 - 3.2. Fingerprinting Methods
 4. Analysis of Population Structure Using Molecular Markers
 5. Molecular Probes for Identification and Characterisation of Marine Phytoplankton
 - 5.1. Introduction
 - 5.2. Probe Design
 - 5.3. Detection Methods
 6. Conclusions
- Bibliography & Internet Resources

Glossary of terms, abbreviations and symbols:

AFLP	Amplified Fragment Length Polymorphism
Biodiversity	A term that can be defined at several scales. Community scale: the sum total of all species in any habitat. Species scale: the sum total of the genetic diversity with each species in a habitat.
DGGE	Denaturing Gradient Gel Electrophoresis.
DNA	Deoxyribonucleic Acid
FISH	Fluorescent <i>in-situ</i> Hybridisation.
ITS	Internal-transcribed Spacer.
LSU	Large subunit of the rRNA.
Molecular Marker	Genetic trait used to identify individuals or sub-populations by its different alleles.
Molecular Probe	Labelled oligonucleotide used to specifically identify a certain taxon of phytoplankton.
Oligonucleotide	Artificially made short piece of DNA.

PCR	Polymerase Chain Reaction. Method to amplify a specific DNA sequence.
RAPD	Random Amplified Polymorphic DNA.
RFLP	Restriction Fragment Length Polymorphism.
rRNA	Ribosomal Ribonucleic Acid.
Phytoplankton	Community of aquatic, unicellular photosynthetic active organisms.
SSCP	Single-strand Conformation Polymorphism.
SSU	Small subunit of the rRNA.
TGGE	Temperature Gradient Gel Electrophoresis.

Summary

Marine organisms are the major, sustaining components of ecosystem processes and are responsible for biogeochemical reactions that drive our climate changes. Despite this, many marine organisms are poorly described and little is known of broad spatial and temporal scale trends in their abundance and distribution. With new molecular and analytical techniques we can advance our knowledge of marine biodiversity at the species level to understand how marine biodiversity supports ecosystem structure, dynamics and resilience. We can then interpret environmental, ecological and evolutionary processes controlling and structuring marine ecosystem biodiversity. With better analytical methods available, we can augment our understanding of biodiversity and ecosystem dynamics in especially the pico- and nano fractions of the plankton as well as in the deep sea benthos, both of which are very difficult to study. We have provided examples of new and long standing molecular tools for researchers in marine ecosystems to enable them to provide better, faster and more accurate estimates of marine biodiversity in the community using tools at the forefront of molecular research.

1. The Importance of Biodiversity Research in the Marine Environment

Understanding and preserving biodiversity was one of the most important global challenges for the past 20 years and will continue to be an important scientific issue into the new millennium. In 1999 the Association of Marine Science Institutes has recognised the need for a science plan for Europe to address the problems associated with a potential loss of biodiversity in the marine environment. This section on the importance of biodiversity research is a synopsis of the executive summary formulated for European research on marine biodiversity reflecting the joint opinions of scientists from the Association of Marine Science Institutes.

The global environment is experiencing rapid and accelerating changes, largely originating from human activity, whether they come from local requirements or from the more dispersed effects of global climate change. Widespread realisation that biodiversity is strongly modified by these changes has generated plans to conserve and protect biodiversity in many parts of the world that were heretofore subject to rampant savaging for natural resources. Adequate ecosystem functioning and therefore the continued use of the goods and services that ecosystems provide to humans depends on how important biodiversity is perceived and preserved. Thus it follows that knowing and recognising biodiversity at all levels is an essential strategy for preserving biodiversity. Basic differences occur between terrestrial and marine ecosystems and the management of their individual biodiversity requires very different approaches. Generalisations concerning biodiversity patterns on both global and regional scales, the mechanisms that determine these patterns, and the consequences of biodiversity loss, are largely extrapolated from the terrestrial ecosystems, and many of these extrapolations are not applicable to the marine environment. Our understanding of marine

biodiversity lags far behind that of terrestrial biodiversity, to such an extent that we do not have enough scientific information to design management plans, such as conservation and the sustainable use of coastal resources. Some of the fundamental differences between marine and terrestrial biodiversity include: the physical environment in the oceans is three dimensional, whereas on land it is only two-dimensional. The main marine primary producers are very small and usually mobile, whereas on land primary producers are large and stationary. Higher level carnivores often play key roles in structuring marine biodiversity and when exploited heavily as in overfishing, there are severe cascading downward effects on biodiversity and on ecosystem functions. This does not apply to terrestrial systems. Marine systems are more open than terrestrial and dispersal of species occurs over much larger ranges than on land. Life has originated in the sea and thus is much older in the sea than on land. As a consequence, the diversity at higher taxonomic levels is much higher in the sea and there are 14 indigenous marine animal phyla, whereas only one phylum is unique to land. The sum total of genetic resources in the sea is therefore inferred to be much more diverse in the sea than on land. Also on average, genetic diversity within a species (i.e. below the species level) is higher in marine than in terrestrial species.

Biodiversity must be evaluated at different scales: These are hierarchical levels (e.g., genetic, species, ecosystems), with spatial scales ranging from single samples to regional and global, and temporal scales changing from short time intervals to long. Threats to marine biodiversity and the consequences of biodiversity loss or change operate at all of these scales. Results from research conducted at any single level can lead to errors and unsupported expectations if extrapolated to other levels. Marine biodiversity is more widely exploited in the sea than on land: man commercialises over 400 species as food stocks from the marine environment, whereas less species are utilised on land. Exploitation of marine biodiversity is also far less regulated than that on land and amounts in the marine environment to the hunting-gathering stage that humans abandoned on land about 10 000 years ago, but technology is becoming so advanced that many marine species are now threatened and many are even extinct.

Marine organisms play pivotal roles in many biogeochemical processes that sustain the biosphere, and provide a variety of goods and services that are essential to mankind's existence, including food production, assimilation of waste and regulation of the global climate. Conservation efforts affect only marine reserves and specially protected areas and the species they contain, which cover at best only a small part of the world's marine water. Thus, adequate functioning of marine systems depends in turn on biodiversity and that fact dictates the need for a broader strategy in the management of biodiversity than conservation alone can accomplish. Any biodiversity project must begin with characterisation of the biodiversity as fully as possible (from genetic to ecosystem level) in selected key (flagstone) habitats across broad geographical ranges. Compiling comprehensive inventories at a few sites should do this and less comprehensive surveys at a larger number of sites, using standardised methods and protocols.

The world's oceans cover 70 percent of the Earth's surface and these areas are dominated numerically by microscopic protists and prokaryotes. The marine phytoplankton are, by definition, high dispersal taxa with large population sizes and are major components of both groups. The bulk of primary production in oceanic and neretic waters involves these small photosynthetic organisms. Until recently most of our knowledge about marine phytoplankton was derived from net samples and bulk process measurements, such as chlorophyll a and ¹⁴C biomass estimates. However, previously unrecognised groups (such as *Prochlorococcus*), size classes (the picoplankton < 3 µm) and hidden biodiversity (new algal classes, e.g., Bolidophyceae and Pelagophyceae) have been found

by utilising whole water samplers and new analytical methods, e.g., flow cytometry, epifluorescence microscopy and HPLC (high pressure liquid chromatography). Surprisingly the picoplankton areas may contribute up to 90 percent of the photosynthetic carbon in certain areas. The picoeukaryotes and *Prochlorococcus/Synechococcus*, whose importance in the open ocean oligotrophic ecosystems has only been discovered within the last 20 years are among this smallest size fraction of the marine phytoplankton.

We may question the accuracy of our knowledge about the genetic diversity of marine phytoplankton with these new revelations into phytoplankton biodiversity. In groups, especially the photosynthetic flagellates, where even α -level taxonomy is lacking, or in groups, such as the picoeukaryotes, where there are far too few morphological markers upon which to determine species identification, we soon realise that we probably know very little about their diversity. In addition we know virtually nothing about the population structure of the phytoplankton. It is likely to be very different from that on land because marine planktonic organisms live in an ever-changing three-dimensional environment. Many taxa may have little genetic structure over very large geographic areas. Further, recent evidence suggests that speciation and dispersal mechanisms in marine planktonic organisms may be very different from those on land. Thus, it is unlikely that generalisations about terrestrial plant diversity and population structure can be extrapolated to marine ecosystems.

2. What Questions Can be Answered Using Molecular Biology Techniques?

The advent of molecular biological techniques has greatly enhanced our ability to analyse all populations, not just the phytoplankton. Their small size and paucity of morphological markers, the inability to bring many into culture, and the difficulty of obtaining samples for long term seasonal studies in open ocean environments has hampered our knowledge of phytoplankton diversity and population structure. Despite this, physiological/biochemical measurements have been used to infer the existence of significant genetic diversity within and between phytoplankton populations. These data have been used to speculate on hidden biodiversity and temporal and spatial structuring of genetic diversity or gene flow. Now molecular techniques can present a quantitative framework through which the diversity, structure and evolution of marine phytoplankton populations can be analysed, predictive models of the dynamics of ocean ecosystems formulated, and the idea of functional groups in the plankton proven.

Molecular analysis of phytoplankton population structure is behind other groups and has been usually inferred from physiological data determined from relatively few clones. This unfortunately is a very naive approach because nearly every physiological measurement has shown that no single clone of any phytoplankton species can be considered truly representative of that species. The need to establish clonal cultures prior to genetic analysis and the inability to perform fine-scale sampling under most conditions are probably the overlying reasons why studies of phytoplankton population structure are perhaps 20 or more years behind those of other organisms. Isozyme analysis, performed for a few species, has revealed heterozygosity between some populations. In addition, fingerprinting analyses, such as RAPDs (random amplified polymorphic DNA) or multi locus probes, have shown that phytoplankton blooms are not clonal but are highly diverse with isolates being related by geographic origin.

The interaction of a species with environmental parameters is influenced by the genetic diversity at the population level of a species. Spatial and temporal partitioning of genetic diversity will occur as

these interactions structure the ecosystem. Such structuring has seldom been measured in the marine planktonic community and studies of genetic diversity are virtually non-existent in pelagic ecosystems. All evidence of geographically isolated populations would be erased if we continue to assume that marine organisms with high dispersal capacities are genetically homogeneous over their entire range. Support for this assumption has come mainly from phenotypic comparisons based initially on net phytoplankton biogeographic studies and later on isozyme studies. Some of the reasons why studies of phytoplankton diversity and population structure have lagged behind those of other organisms is because of their small size and the lack of morphological markers, and the ability to bring into culture only a small part of the known biodiversity. The lack of knowledge of their breeding systems makes genetic or demographic studies difficult. Logistical problems of collecting samples for long term seasonal studies in open ocean environments or for doing fine-scale sampling are additional reasons.

Another issue is whether adequate sampling strategies can be employed for phytoplankton populations to address spatial and/or temporal genetic variation questions. Pre-established cruise tracks may make the sampling of oceanic populations only possible at depth rather than in a hierarchical grid-like fashion that may be needed for population studies. A lack of knowledge about current regimes in the study area may also bias sampling strategies if samples are unknowingly taken from two water masses. At present most genetic studies must rely on clonal cultures for their analyses. These single-cell isolations are made from natural populations and can be difficult to perform at sea. The selective survival of only 10 – 30 percent of clones from natural populations may mean that the range of genetic diversity determined from a bank of clonal isolates may not be a true reflection of the genetic diversity in the original population and may not be adequate for the level of genetic diversity being addressed. In many algal groups, life histories are incomplete, and if the algae undergo sexual reproduction during culturing, then this may also affect the type of genetic analysis performed

As far back as the mid-seventies Doyle hypothesised that planktonic algae must consist of a multitude of competing genotypes, but this study was largely ignored and it was assumed that planktonic taxa may have little genetic structure over very large geographic areas. Marine planktonic organisms live in an ever-changing three-dimensional environment and it was assumed that highly dispersed organisms at the mercy of the currents have no trace of genetic structure.

With the advent of nucleic acid methods, however, these views on the absence of genetic structure in the marine phytoplankton have been seriously challenged. Genetic structure and physical, spatial partitioning within biogeographic regions are now known. The idea of a single globally distributed species is no longer believed, nor is the idea of temporal stasis. Temporal genetic changes can often be greater than spatial changes or changes between species. This may very well apply to bloom populations. The rate of genetic change can and does occur on ecological time scales. Reasons for this are unclear but such changes may play a role in determining how local adaptations and speciation can occur in apparently homogeneous populations. The concept of a 'super species' with the ability to exploit a wide spectrum of environmental conditions may lay the groundwork for temporal genetic change.

Much of our limited knowledge about phytoplankton genetic diversity stems from the difficulty of finding polymorphic markers for ecological genetic studies. Isozymes, the molecular genetic markers used in early studies, evolve so slowly that closely related populations appear identical. This fact has undoubtedly propagated the early ideas of the absence of genetic diversity in marine

phytoplankton. The use of high resolution DNA fingerprinting techniques *sensu lato* circumvents these problems and has thus opened areas previously considered intractable.

Plastid and flagellar apparatus characteristics are the features that define most phytoplankton classes, making them monophyletic taxa, but some surprises have been revealed by molecular analyses. For example, the Euglenophyceae, once thought to be related to the Chlorophyceae, are shown to be a very early eukaryotic radiation and not part of the major eukaryotic radiation called the crown group radiation. The kingdom Chromista did contain the bulk of marine eukaryotic phytoplankton taxa, e.g., the Heterokonta, Haptophyta, and Cryptophyta. But this kingdom is now recognised as a polyphyletic taxon. Molecular analyses based on total evidence (both morphological and molecular data from the rRNA data set) continue to reinforce the clear separation of the haptophyte from the heterokont algae whereas those based on many other genes have distanced the cryptophytes from both the heterokonts and the haptophytes. A fourth group, the Chlorarachniophytes, were also formerly placed in the Chromista but are now shown to be clearly related to the foliose amoeba. Clearly the Kingdom Chromista is an idea whose time has past. New algal classes have been recognised from molecular analyses, e.g., Pelagophyceae and the Bolidophyceae.

Molecular techniques have changed systematics dramatically at the genus and species level, showing polyphyletic and paraphyletic lineages across many algal groups, not just the phytoplankton. The most dramatic upheavals have come in groups with few morphological markers, and where morphological species definitions have been too broad. The prochlorophytes, *Chlorella*, *Chlamydomonas* and *Chrysochromulina* are all recognised as polyphyletic taxa. But even in groups with good morphological markers, e.g., *Skeletonema* and *Cryptocodinium*, (cryptic) sibling species; have been found; others not so easy to differentiate at the species level, e.g., *Phaeocystis* also contain cryptic species.

Molecular tools in general offer the possibility to estimate biodiversity at all levels, e.g., kingdom/class/family/species level, in a comparatively small environmental sample. In some cases even a few milliliters of seawater may be enough. Moreover, some of the techniques are very sensitive, e.g., offer the possibility to detect single cells in a sample. Depending on the question(s) being asked the molecular tools to answer them differ greatly. One may wish to detect as many species as possible in a given sample. In this case the establishment of an rRNA clone library with subsequent sequencing of as many clones as possible can uncover the biodiversity in the sample in great detail. General assessment of comparative biodiversity in a larger number of samples can be achieved with fingerprinting methods based on restriction fragment length polymorphisms (RFLPs), denaturing or temperature gradient gel electrophoresis (DGGE, TGGE) or single strand conformation polymorphisms (SSCP). Presence or absence of a known species can be monitored with species-specific probes using chemiluminescent detection with dot blot techniques or, more sophisticated, with fluorescent *in-situ* hybridisation (FISH). Distinction of individuals at the family or even species level can be obtained using highly variable molecular markers such as ITS sequences (inter-transcribed spacer) or microsatellites. Finally per se non-molecular techniques like flow cytometry that have already been used in the “pre-molecular age” can be combined with DNA techniques (staining of the nucleus, hybridisation with fluorescence-labelled specific oligonucleotide probes) to distinguish and quantify species in environmental samples.

In general molecular techniques have some significant advantages over traditional methods:

1. Only very small samples (in the range of milliliters up to a liter) are required for most analyses.

2. Sensitivity of many methods is very high, e.g., enabling the researcher to detect even single specific cells among thousands of others.
3. Dead or non-culturable cells can be analysed.
4. Species-specific data (such as sequences) can be obtained without the need to culture or even isolate a species.

As with all methods, molecular ones also contain certain biases. The harvesting of cells through filtration or centrifugation may be harmful for fragile organisms, which thus may escape the analysis. For many techniques the lysis of organisms with subsequent isolation of DNA is a prerequisite. Both steps may not be equally effective in all organisms. In PCR-based approaches biases are evident concerning the choice of (universal) primers, PCR conditions (e.g., the amount of DNA or primers used, the annealing temp., cycle number etc.), machines or enzymes used etc. The copy number of genes of interest (mostly ribosomal RNA genes) differ greatly among various organisms. If cloning steps are involved, then the choice of vectors, enzymes or bacterial strains may be relevant. Hybridisation experiments are susceptible to hybridisation conditions (temperature, salt concentration, time) or base composition and subsequent detection of fluorescence may be hampered by autofluorescence. All the former is especially important when absolute quantification of results is desired. In general we advise the same caution when interpreting the results of molecular methods as for all other methods. Results are not more reliable because they come from a “molecular” approach rather than a “classical” one.

In the following we will summarise and briefly explain a variety of techniques currently being used or under development. We also try to estimate the advantages and shortcomings of such methods.

3. Evaluating Marine Biodiversity by Sequence Analysis and Fingerprinting Methods

3.1. Sequence Analysis

3.1.1. Which Genes to Select?

The phytoplankton, being pigmented organisms, contain three different genomes: the nuclear, the plastid and the mitochondrial genomes. Each has their own unique set of genes, which evolve at different rates. To use molecular techniques in phytoplankton analyses, one must be aware of certain biases and limits to the resolution of the genes. Researchers working with photosynthetic organisms have a different array of genomes/genes to access than do those working with heterotrophic organisms. However, many genes are shared by all organisms irrespective of their nutritional class, and these genes should be used if one needs to carry out very broad comparisons. Several questions should be considered when selecting a molecular marker for phylogenetic or population structure analyses. Answers to these will strongly influence the molecular markers selected for a study.

However overriding any consideration of which region of the genome to select will be two other factors: Is the rate of evolution in the chosen molecular marker appropriate for the taxonomic level addressed and what is the geological age of the species or group of organisms investigated? If the algal group or any group, for that matter, is ancient and the rate of evolution in the chosen molecular marker fast, then the molecular marker may be so saturated with substitutions that no phylogenetic signal can be recovered. If not saturated, then population level variation or cryptic species may be detected. In more recently evolved taxa, slower evolving genomic regions will lack resolution, and non-coding regions may even be inappropriate at higher taxonomic levels.

At higher taxonomic levels, slower evolving genomic regions, e.g., coding regions, such as the ribosomal RNA genes (small subunit or SSU, large subunit or LSU) and the large subunit of RUBISCO (ribulose 1,5-bisphosphate carboxylase, gene name: *rbcL*), are commonly used. The use of other genes, such as the *tufA* gene, is increasing with the advent of the polymerase chain reaction. By far the most sequences are available for the SSU rRNA. This molecule, besides the availability of a huge dataset, offers some advantages:

1. It is present in all organisms and organelles capable of protein biosynthesis because it is part of the ribosome.
2. Most parts of the gene have been conserved to a high degree in the course of evolution. Nevertheless, it can be difficult to align. Also primers for the amplification of the molecule via PCR can be designed without problems.
3. The SSU gene mostly is present in many identical copies (sometimes in the range of tens of thousands!) in the genome and is thus easily PCR-amplifiable even from very small amounts or degraded DNA than other genes.
4. For many SSU molecules the secondary structure is known, making it easy to gain insight into the function of certain areas of the molecule.
5. The fact that the molecule *in vivo* shows a significant secondary structure makes it the first choice for fingerprinting techniques, e.g., SSCP, which are based on secondary structure formation.

One should be aware, however, also of some disadvantages of the SSU for phylogenetic reconstruction. Because it is a non-protein coding gene its sequence can be only aligned at the nucleotide sequence level. Thus, only four variables (the nucleotides A, C, G and T) are present as compared to the amino acid sequence of a gene (20 different amino acids), it can be aligned less effectively, especially in variable regions. Because of its high degree of conservation (see above) resolution in phylogenetic trees is sometimes poor, especially between closely related species. Because of the limited sizes of the molecule (e.g., 1500 – 1800 bases), in combination with its overall small variability, it sometimes does not contain enough informative sites to answer fundamental evolutionary questions. Nearly all comments about the SSU also hold true for the large subunit rRNA (LSU), except that this molecule is significantly larger and contains more variable sites.

Despite the advantages of rRNA genes for some applications the use of protein-coding genes may be more appropriate. For all photosynthetic, and also some chemosynthetic, organisms the gene (*rbcL*) for the large subunit of the key enzyme of the Calvin cycle, ribulose-1,5-bisphosphate carboxylase/oxygenase (RUBISCO) has become a keystone in phylogenetic comparisons. The gene is rather large (approx. 1600 bp) and well conserved, although to a lesser degree than the SSU. Because it is a protein-coding gene it can be unambiguously aligned on the amino acid level. The *tufA* gene encodes a translation elongation factor, is present in all organisms, like the SSU gene and is well conserved and easy to align. As the dataset for many genes grows, others, e.g., RNA polymerases, GAPDH, COX1, might also gain importance in the future.

At lower taxonomic levels, non-coding spacer regions may be more appropriate, because these are commonly far less conserved than coding regions. The best-studied examples are the internal transcribed spacer regions (ITS) within the ribosomal cistron, which are those regions separating the SSU and LSU genes. Sometimes also external spacers of non-transcribed spaces can offer an even greater variability than the ITS regions or introns. The spacer separating the large and small subunits of RUBISCO (only in the chlorophyll-c and the red algae but not in the dinoflagellates), the spacer between the pet B and D genes, the spacer between the trnT and the trnF genes and the introns in

the calmodulin genes are commonly used. These regions are not under the same functional constraints as those of coding regions and therefore are free to evolve at a faster rate than coding regions and can provide greater resolution among closely related species.

3.1.2. How to Generate Sequence Data?

The starting point for this task is the availability of a pure culture or isolate. (Examples of how to start with mixed populations are given in section 3.1.3.). From any sample, nucleic acids have to be isolated. Generally this would be DNA but sometimes there is the need to start from RNA. For this isolation many commercial kits are available but classical methods, such as phenol or CTAB extraction, often work as well. (In some cases even crude extracts, e.g., resulting from just boiling the sample and taking the supernatant, can be used for PCR amplification.) Usually from this DNA the gene of interest is then amplified using PCR. For all genes listed above, as well as many others, universal primers are available. The PCR product is then either cloned and sequenced, or sequenced directly. There is a growing number of companies that offers low-cost sequencing services. In general cloned fragments are easier to sequence than PCR products. With modern cloning kits, such as the TOPO vectors, cloning is easy to perform, nevertheless the cloning procedure costs extra time and money.

In some cases it may be of interest not just to determine a sequence of a gene of interest but also to decide whether the species it was isolated from was physiologically active at the time of sampling. Here one would isolate RNA rather than DNA, back translate the RNA into cDNA via a reverse transcriptase (RT) step and then amplify by PCR the gene from the cDNA. The entire process is called RT-PCR and even allows to estimate semiquantitatively the amount of specific mRNA present in the cell. The underlying assumption is that only in active cells are genes transcribed into mRNA that then can produce a PCR product. The necessary control experiment to perform is to also PCR amplify the RNA without the RT step. This is to show that the RNA is not contaminated with DNA.

3.1.3. Determining Biodiversity in an Environmental Sample by Sequence Analysis

The most exact method to assess biodiversity down to the species level in environmental samples is by the determination of sequences of clones from such samples. The SSU rRNA gene is often the gene of choice for cloning being that most commonly used as a phylogenetic yardstick. This is best achieved by isolating total DNA from the sample and then full-length SSU gene amplification using PCR and universal primers. The resulting PCR products represent the diversity of organisms present in the sample. PCR products are then cloned resulting in a clone library of the sample (Figure 1). If necessary redundant (identical) clones are identified and the library is randomly sequenced. Based on partial sequences the clones can be phylogenetically classified and the complete sequence of clones of interest can be determined. Technically it is no problem to generate more than 5000 clones from a single PCR reaction, therefore saturation of clone libraries can be assumed. The method thus allows the exhaustive description of biodiversity in a sample down to the species level. Also the resulting sequence information may serve as a basis for developing specific oligonucleotide probes necessary for subsequent methods like FISH. Disadvantages of the method are:

1. It is generally very labour-intensive, especially when redundant clones are identified and removed. This can only be done by comparing individual clones by digestion of the insert after performing

plasmid isolation or a second PCR reaction. In the future the use of robots for plasmid isolation may help solve this problem.

2. Sequencing a large number of clones though easily performed with modern sequencers is still costly. So the method is not yet suitable for screening of large numbers of samples. However, as prices for sequencing services drop it may be better usable in the future. The method may already serve to establish an initial data bank for a given sample area.

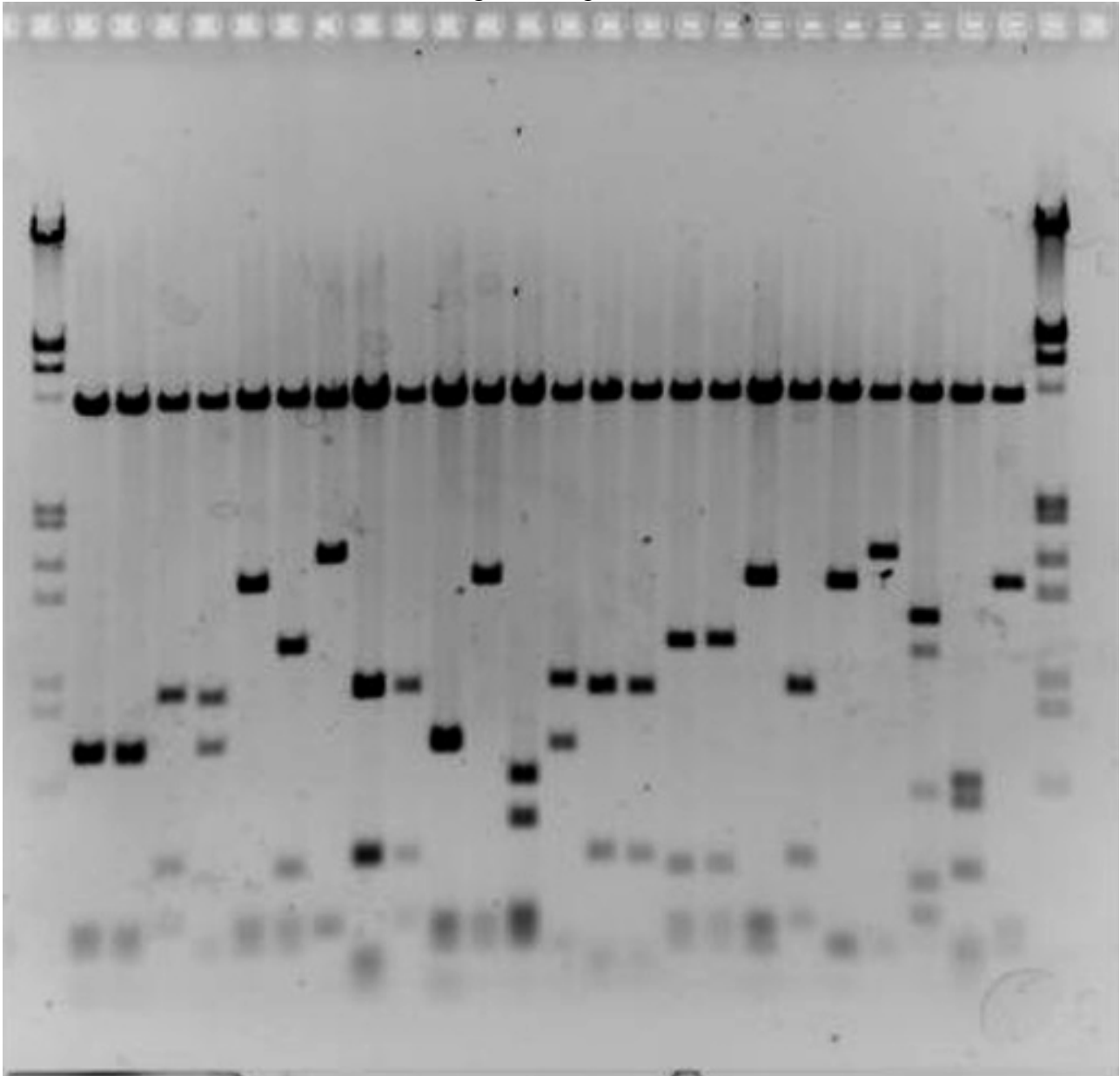


Figure 1: Biodiversity of marine Picoplankton. DNA was isolated from an environmental sample, i.e., 3 µm filtered sea water. A clone library of PCR-amplified SSU rRNA was established and plasmids were analysed by restriction enzyme digestion. The figure shows the typical variability of such clone libraries. (Lanes 1 & 26 = size markers, lanes 2-25 = 24 individual clones digested with restriction enzymes.)

3.1.4. Analysing Sequences for Determining Phylogenies and Biodiversity

Once a sequence dataset is established there is the need to analyse it further. As a first step, sequences can be compared to existing databanks (GenBank, RDP, EMBL etc.), in most cases by performing a "BLAST" search (Basic Local Alignment Search Tool). This can be done on-line via the

internet. As a result it is possible to tell quickly whether or not the determined sequence is identical to other counterparts in the databanks. If not, the phylogenetic affiliation of the sequence can be estimated, easily to the class level, sometimes down to the family level. In most instances sequences need to be aligned to (sometimes large numbers of) other sequences. This can either be done by constructing a completely new alignment or by adding new sequences to an existing one. The latter is more appropriate when a comparably small number of new sequences is added to a large alignment, the former if the number of new sequences exceeds the number of sequences in an existing alignment. The public alignment package CLUSTAL produces good alignments from scratch but other programs may be used as well. Many programs can be used online thus eliminating the need of big in-house computer power. Adding new sequences to large existing alignments can very well be done with the ARB package. This program is also strongly suggested for further applications like tree construction or especially the design of specific oligonucleotides probes. Generating an alignment is the prerequisite for further applications like phylogenetic reconstruction or oligonucleotide probe design. Therefore care should be taken in constructing it. If possible it should be checked manually. Any further information, e.g., on the secondary structure of the aligned molecule, should be included.

Once an alignment is constructed, phylogenetic relationships of sequences in the alignment can be analysed. Phylogenetic trees from such analyses can be interpreted better if an outgroup organism is included in the alignment. Ideally such an organism is closely related to the systematic group under investigation but stands outside of it. This organism then can be placed at the bottom of the tree, thus "rooting" it. For example, choanoflagellates or sponges root the animal kingdom, being the most primitive of the Metazoa.

For phylogenetic reconstruction three major methods during the last decade turned out to be most appropriate: Parsimony, neighbour-joining, and maximum likelihood. Parsimony counts for every given pair of sequences in the alignment the minimum number of mutational events to transform one sequence into the other. The tree is then constructed in a way the number of such "parsimonious" steps in the entire tree is minimal. Despite the fact the underlying assumption of the method (e.g., that evolution marches forward unidirectional in the fastest possible way) is most likely wrong, parsimony trees seem to produce a good picture of evolution. A disadvantage of the method is that the computing power needed for trees containing more than 10-12 sequences quickly exceeds what is currently present on earth. In this case not all possible trees can be analysed bearing the risk the "best", e.g., most parsimonious, tree is not found. However, there are methods to minimise this problem. Parsimony analyses can be performed with PAUP, PHYLIP, ARB, and a number of other computer programs.

Neighbour joining (NJ) first constructs a distance matrix for the alignment by calculating distance scores for every given pair of sequences in the alignment. For this purpose a number of matrices are available that take into account different exchange rates observed for different residues at a given position in an alignment. By this e.g., transitions can be weighted differently than transversions (nucleotide level) or conservative amino acid exchanges can be counted less than non-conservative exchanges. Then the tree is constructed by first grouping most similar sequences and subsequently adding less similar sequences or groups of sequences. Thus, a tree is constructed whose summed-up overall distances are minimal. The NJ algorithm does not require much computing power and thus, tree building is very fast. Therefore many "bootstrap cycles" can be performed. An easy-to-use NJ program is TREECON but many others are available.

Maximum Likelihood (ML) is a very computing-intensive tree building algorithm that compares the likeliness of a given grouping in the tree with all other possible clusterings, thus producing the "most likely" tree. Many researchers believe that this is the "best" available method but this believe is difficult to prove. ML can - among others - be performed with PHYLIP or with ARB.

Reliability of trees generated by all methods listed above can be judged by "bootstrapping". In this approach from a given alignment of n positions, n columns are chosen randomly and added to each other until a new artificial alignment of identical size is generated. This new alignment is thus based on the original one but highlights certain areas of it and under-represents others. This procedure is repeated many times and for each of the alignments, trees are constructed. It is then counted how often a certain clustering re-appears in all trees and this number is written at this branch as the bootstrap value. It is difficult to judge the significance of bootstrap values because aspects like the size of the alignment, the number of species in the tree, or the overall degree of conservation in the molecule may have an influence on this value. In our experience values below 70 percent suggest not to trust a given branching order, and values above 70 percent may be regarded as moderate to strong evidence for correct clustering.

Perhaps the best thing one can do to reconstruct phylogenetic relationships is to use all three methods listed above and do bootstrapping on all three trees. Then only those branches should be given that are supported by at least 70 percent bootstrap values and all the others should be drawn as unresolved.

Molecular evidence can also provide an objective framework with which to reconstruct the historical biogeographic distribution of taxa as well as recover recent dispersal events and to estimate divergence times. By using the fossil record or geological events to date divergences between species, it is possible to correlate present-day distributions of taxa with their biogeographic history and to date nodes in the tree where no fossil date is known. If there is no fossil record, but the phylogenetic reconstruction of a group is congruent with a biogeographic history of the study area, then it is possible to correlate the divergence of taxa with palaeo-oceanographic and palaeo-climatic events to explain their present-day biogeographic distribution. Phylogeography, a new field, has been termed for such studies. For marine plankton communities, dates for the opening and closing of oceans, for the movements of the continents relative to the water masses and for climate changes resulting in fluctuations in sea-level are used, unless the group has a detailed fossil record, like the diatoms, coccolithophorids and the dinoflagellates for dating divergences. It is also important to calculate relative rates of evolution prior to estimating divergence times to ensure that fast-evolving species, which could bias the determination of divergence times, are eliminated from such calculations.

3.2. Fingerprinting Methods

Two well-established methods for assessing diversity in environmental samples are Temperature Gradient Gel Electrophoresis (TGGE) and Denaturing Gradient Gel Electrophoresis (DGGE). These methods allow the qualitative and semi-quantitative determination of biodiversity in environmental samples. Total DNA from such samples is isolated and specific DNA fragments are amplified by PCR. Usually the gene for the small subunit ribosomal RNA is chosen for this purpose and a fragment of approx. 500 bp length is amplified using universal primers, e.g., primers that recognise a wide variety of species (e.g., all eukaryotes, all prokaryotes or Archaea). Of course also class/family etc.-specific primers can be used. Because the high degree of conservation in the SSU gene PCR products will all have approximately the same size, but they more likely will differ in GC

content and in the amount of possible secondary structure formation. As a result individual PCR products will denature to different extents when treated with denaturing conditions like urea, formamide (DGGE), or elevated temperature (TGGE). (Complete denaturing is prevented by the use of a “GC clamp” in the primers, e.g., a GC-rich stretch that will not denature under conditions used.) Therefore individual PCR products can be separated by electrophoresis in polyacrylamide gels containing either a gradient of denaturing chemicals (DGGE) or that are run in a temperature gradient apparatus (TGGE). A typical environmental sample in this assay will produce a large number (e.g., > 50) of bands and many samples can be compared on a single gel. Within limitations band intensities correlate with the relative abundance of species in the sample producing them. It is possible to excise individual bands from the gel, reamplify or clone them and determine their sequence thus allowing to assign species to certain bands. Based on the sequence information specific oligonucleotides can be designed that recognise the previously sequenced nucleic acid in FISH experiments. Limitations of DGGE/TGGE are:

1. Due to the separation capacity of polyacrylamide gels the PCR fragment size is limited to some 500 bp. Therefore the methods can not handle full-length SSU genes. Based on results of DGGE/TGGE it is thus difficult to obtain full-length SSU sequences from species of interest.
2. It is difficult, if not impossible, to compare patterns across gels. Therefore the number of samples that can be compared with each other is limited to the number of slots on the gel.
3. The methods are not trivial in handling and have to be optimised for a given primer pair and for new sample types. So at least the initial help of an experienced person is necessary to establish the methods. However, given the quickly accumulating data set on this method in the literature this problem may become less relevant in the future.
4. Sometimes the methods are “too sensitive” in that even pure cultures produce more than one band.

Similar to DGGE and TGGE single-stranded-conformation polymorphism (SSCP) uses the fact that single stranded rDNA fragments fold into secondary structures depending on their base composition. In contrast to the above mentioned methods, in SSCP single-stranded rDNA fragments without GC-clamps are used. Based on their folding fragments of identical size, but different base composition can then be separated in non-denaturing polyacrylamide gels. Single-stranded rDNA fragments are generated as follows: A fragment (e.g., of 300 to 500 bp length) of the SSU rRNA gene is amplified by PCR. In this reaction the forward primer is dephosphorylated and the reverse primer phosphorylated. After amplification the PCR products are digested with Lambda Exonuclease which specifically degrades the phosphorylated strand and, thus, removes the non-coding DNA strand. SSCP of DNA from environmental samples typically result in a complex band pattern comparable to DGGE or TGGE. Bands can be excised from the gel, reamplified by PCR, cloned or sequenced, and gels can also be blotted. As with DGGE/TGGE in principle the method works with all sequences that can form secondary structures. For some questions the V4 & V5 region of the SSU rRNA has proven most useful. Also, instead of universal primers specific primers, e.g., for actinomycetes or fungi can be used. The major advantage of SSCP compared to DGGE or TGGE is that (i) no GC-clamp needs to be generated and the electrophoresis method is more straight forward, because ordinary equipment for PAGE (polyacrylamide gel electrophoresis) instead of gradient gels or temperature gradient electrophoresis can be used. Also the fact that one of the two strands is degraded reduces the variability obtained from communities because it avoids heteroduplex formation, a problem known in community analyses based on DGGE. A current disadvantage of SSCP relates to its novelty and thus, lack of applications in microbial ecology. However, because also automation is possible SSCP may become more important during the next years.

The shortcoming of the above-mentioned methods, e.g., the size limitation of separable fragments, may in the future be overcome by the use of sequence specific dyes in agarose gel electrophoresis. This approach is based on the availability of sequence specific intercalating dyes that are linked to a PEG molecule (polyethylene glycol). Double-stranded DNA fragments of identical size, but different base composition, treated with such dyes bind them in a sequence-specific manner resulting in different mobilities in agarose gels. Given that it occurs at the dye binding site, a single point mutation in a 500 bp fragment is detectable with this method. The feasibility of agarose gels makes it possible to analyse larger fragments than in DGGE, TGGE or SSCP. We could clearly separate fragments in the range of 500 to 1200 bp and it was possible to distinguish from each other full-length SSU PCR products from several axenic algal cultures. Until now, however, only two such dyes are commercially available and it may be due to this fact, that separation of larger fragments is yet rather poor.

4. Analysis of Population Structure Using Molecular Markers

Whereas fingerprinting techniques like DGGE etc. find their use in analysing samples of unknown composition, e.g., field samples with a mixture of species, the molecular markers we describe in this chapter are used to investigate diversity at the level of individuals or strains/sub-species of one known species. A molecular marker as we understand it is a genetic trait of DNA or in one case protein (isozymes) used to distinguish between individuals or groups by the marker's different alleles. There is no such thing as the perfect molecular marker and every technique has its advantages and disadvantages and must be chosen considering the focus of: the scientific question to be solved, the species under investigation and one's previous knowledge about the species as well as available resources (time, equipment, costs).

The first molecular markers to be used in the marine field, as in the terrestrial area, were the isozymes. These are proteins that show only small differences in their size or iso-electric point and therefore can be separated by electrophoresis but are still able to catalyse the same biochemical reaction. Their advantages of quick and easy isolation and detection made them the markers of choice for many investigations. But the requirement that isozymes must still be functional in the biochemical pathways strongly limits the number of possible mutations and therefore the number of alleles and the heterozygosity of this marker type. Another disadvantage of this kind of marker is also that protein content of cells and following the detectability of isozymes is strongly influenced by the environment and as a consequence, marker types were developed that directly used environment-independent DNA.

DNA polymorphisms between individuals could, e.g., be found by Restriction Fragment Length Polymorphism (RFLP), a technique in which DNA is digested by restriction enzymes and then one compares the presence or absence of restriction sites in different individuals as well as insertions or deletions in their genome if they lack these restriction sites. The "classical" way of detecting this kind of polymorphism is by a Southern blot hybridisation with single copy genes as probes. The normally radioactive-labelled probes bind to a DNA fragment and are detected by autoradiography. If these fragments are of different length in two individuals because of the mutation processes mentioned before, then two different sized bands, respectively alleles, can be seen on the x-ray film. This method also made it possible to distinguish between homozygote and heterozygote individuals (co-dominant markers) and has therefore a higher information content as a dominant marker. But one can also see that this technique is very laborious and time-consuming when testing all different kinds

of restriction enzyme / probe combinations to find polymorphic bands and also needs large quantities of high-quality DNA for digestion and blotting. It is also hardly possible to automate this technique which makes it difficult to process large number of samples.

A slightly different RFLP method consists of the PCR amplification of a specific gene, e.g., the SSU rRNA sequences followed by restriction digestion with enzyme and gel electrophoresis. Because it uses only a limited number of fragments this method avoids the need of blotting and probing for visualisation and is much faster and easier than the "classical" RFLP. But the limited number of possible bands leads also to a very small number of possible polymorphisms and one needs luck to find a usable marker. Nevertheless, there are examples where this kind of RFLP marker has been used with success, e.g., for discriminating species and strains of the toxic dinoflagellate genus *Alexandrium*.

The first widely used PCR marker technique was Random Amplified Polymorphic DNA, (RAPD) or Arbitrary Primed PCR (AP-PCR) with the former being the most commonly used name for this kind of method. It uses a single short random primer in a PCR reaction, mostly decamers, to amplify the DNA between the primers to give a fingerprint of multiple bands and polymorphisms between individual samples are derived from single nucleotide changes that prevent or allow primer binding and therefore lead to different banding patterns between individuals. This method became quite popular because it could be carried out in a short time without previous knowledge of the organism under investigation. Most marker-assisted phylogenetic studies in the marine field published so far, e.g., for the prymnesiophyte *Emiliania huxleyi* that plays a major role in the ocean's carbon cycle or for the toxic dinoflagellate *Gymnodinium catenatum*, have used this technique. Nevertheless, RAPDs have been shown to have some drawbacks: The use of short primers gives not only the possibility of random binding in all kind of genomes and therefore makes this method working at all, but it also makes it unreliable, too, and susceptible even to small changes in the PCR conditions and manpower. As a consequence reproducibility of RAPD markers is hard to obtain.. Also, RAPDs are normally dominant markers by which they give less information than other, mostly co-dominant markers. To summarise, RAPDs are only the method of choice when time and resources are limited and no previous information about the species under investigation are known, else other marker techniques are to be preferred.

A more recent marker technique for studying biodiversity in the marine environment is called AFLP (Amplified Fragment Length Polymorphism), which combines the advantages of RAPDs and RFLPs into a powerful tool. First, genomic DNA is digested by two different restriction enzymes, a rare and a frequent cutter, then matching adapters are ligated to the digested fragments. Afterwards, a PCR is performed with primers homologous to the adapters plus up to four additional random bases at its 3' end. By using these selective bases, only a subset of digested DNA fragments is amplified, giving distinct bands instead of a smear and making it possible to analyse the bands on a polyacrylamide gel.

The major advantage of this technique is the large number of bands it produces, giving a very good chance of finding a large number of polymorphic bands among them. The polymorphisms detected by this method come from the same sources as in RFLPs, insertions, deletions and point mutations leading to the presence or absence of restriction sites, but compared to RFLPs, AFLPs are normally scored only as dominant markers, even when some researchers gave possible methods for using them co-dominantly.

The use of longer PCR primers that anneal to the adapters and a few bases of the genomic DNA make the whole reaction much more reliable than RAPDs, because higher annealing temperatures can be used. The greatest advantage of the RAPD technology on the other hand remains, because no previous sequence information of the species under investigation is needed and PCR reactions are fast to be carried out. Nevertheless, AFLPs are technically demanding, sensitive to the purity and quantity of DNA to be digested, need some experience to be performed and data analysis of the hundreds of amplified bands must be done by computer analysis. Since 1995 when AFLPs were first introduced, there has been an increasing number of publications using this technique, but most of them deal with higher plants either for developing genetic linkage maps or for doing population studies. In algae only two publications have so far emerged, both dealing with the multicellular red alga *Chondrus crispus*, but more will definitely follow, including those for phytoplanktonic, unicellular alga (Figure 2).

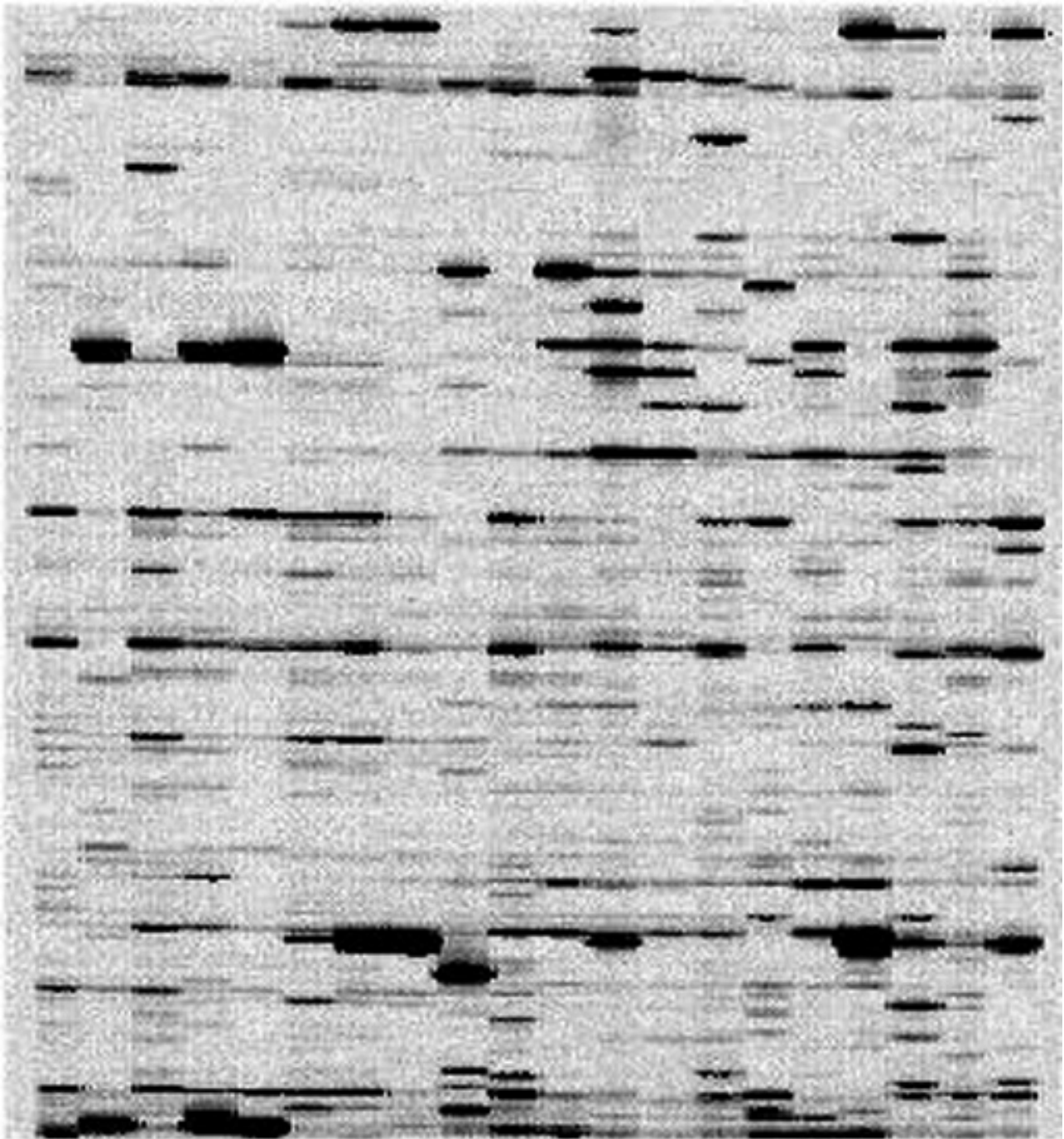


Figure 2: The fingerprinting method of Amplified Fragment Length Polymorphism (AFLP) shows clear differences between different isolates from one population of the toxic

dinoflagellate *Alexandrium tamarense* isolated from the Orkney Islands, Scotland. (Photo courtesy of U. John, unpublished)

Microsatellites, also called simple sequence repeats (SSR), are already widely used as molecular markers both in applied genetics and in studying biodiversity in the marine field. Examples are mainly from the field of fisheries sciences with most if not all economically important species covered but some first results were also published for macroalgae (*Gracilaria gracilis*, *Laminaria digitata*) and microalgae (*Chlamydomonas reinhardtii*, *Emiliania huxleyi*, *Ditylum brightwellii*). Microsatellites are short repeats of one to six nucleotides, e.g., (CT)_n or (CAG)_n, repeating themselves five to dozens and sometimes hundreds of times. They are found in great abundance dispersed all over the genome of all organisms investigated so far. This abundance together with the large number of alleles resulting from high mutation rates because of their special, regular structure make them highly useful molecular markers at the population level. Microsatellite polymorphisms can be revealed where other marker types have failed and therefore they are especially useful for species that otherwise lack a high degree of polymorphism, such as inbreeding species like important crops as soybean, or clonal species like planktonic algae that don't have a regular sexual cycle. Comparisons of different marker types have shown that microsatellites have the highest degree of polymorphism of all commonly used marker types.

Using microsatellites as markers is easily done by PCR reactions with primers homologous to the SSR flanking unique sequences and analysing the amplification products by electrophoresis on high-resolution gels (e.g., Polyacrylamide or MetaPhorTM agarose). Different alleles can be detected that way, because mutation events in microsatellite sequences are normally changes in length by errors during replication or by unequal cross-overs. Identification of microsatellite alleles by their size also means that they can be scored as co-dominant markers with all their advantages in population studies. Compared to the ease of using existing microsatellite markers, the biggest drawback of this technique is the often time-consuming way to develop the markers in the beginning. Also, in most cases the transfer of microsatellite markers between different species is not very successful, which means that new markers have to be developed for every new species under investigation. The easiest way to do this is to screen databases like EMBL or GenBank for microsatellites in already published sequences of one's species of interest and to develop flanking primer pairs directly from that. Of course, this is only possible if one is working on a "popular" species for which a certain number of sequences already exist. The other, more time-consuming way is to screen a genomic library of the target species using microsatellites containing oligonucleotide probes, isolate and sequence positive clones and design the primers from that source. This costly method of course makes only sense, when other methods fail to give the resolution needed, and the cost factor may also be the reason for the lower number of publications using microsatellites in population studies in comparison to AFLPs. The application microsatellites are most used for so far is the development of genetic linkage maps as it is done for all major crops to help in breeding research, but this will change with time when more microsatellite markers or sequence information, e.g., from genome projects, are available and population studies using this marker type will be done more often.

For analysing populations on the level of the individual or of closely related strains we suggest to use either AFLPs or microsatellites. Both techniques have been shown to be the cutting edge of marker technology at the moment with AFLPs giving many scorable bands with a good chance of polymorphic ones and its usability without the need for previous knowledge on one hand, and with the easier-to-use microsatellites with their normally extremely high degree of polymorphism that

exceeds every other marker technique so far but it has high set up costs that must be taken into account.

Another thing the user should be aware of is that the type of marker also influences the kind of algorithm that has to be used for analysing the resulting data. Microsatellite polymorphisms occur due to stepwise changes in the number of repeats and this leads to another algorithm for calculating genetic distances than, e.g., the insertions, deletions and point mutations that can produce different RFLP alleles. Also, it is important for data analysis of diploid or polyploid organisms, if the marker is co-dominant or only dominant, making it impossible to distinguish between heterozygotes and homozygotes which results in less informativeness of the data. In a substantial number of phytoplankton species the level of ploidy is not yet known or may change during different stages of the cell's life cycle. Together with sometimes morphologically indistinguishable cell types in these life cycles, it should be clear that the data analysis of molecular markers from marine microbial populations is not a trivial task and should be performed with great care.

5. Molecular Probes for Identification and Characterisation of Marine Phytoplankton

5.1. Introduction

Quite often morphological features as seen by light microscopy are not sufficient to distinguish clearly between species or groups of phytoplankton or marine bacteria. Therefore, more sophisticated methods like electron-microscopy or analysis of specific chemical components by HPLC are needed to identify a species for sure, but these are laborious and time-consuming. An alternative approach is the development of specific molecular probes. These probes are short oligonucleotides of normally 16-24 bp length that are hundred percent homologous only to a complementary sequence in a gene of the species of interest and differ by at least one position to all other organisms. In hybridisation experiments these probes can therefore be used to identify species of interest by binding to the target's sequence and later detection by a probe-attached label, e.g., Digoxigenin (DIG) or a fluorochrome like fluorescein.

The application range of these probes extends from answering ecological questions like species composition and its change through space and time to the development of an early warning system for harmful algal blooms using probes for toxic species.

Probes were normally developed from genes for which a larger number of sequences from different organisms exist, such as, the small and large ribosomal subunits and their ITS regions, that have been widely used in phylogenetic studies. The reason for this is the possibility to compare the possible probe to a broad data set to see if it is really specific. The use of sequences of the rDNA has also other advantages for probe design. First, this molecule has regions with different degrees of conservation, which makes it possible to develop probes for higher taxonomic groups (class level probes, e.g., for prymnesiophytes), probes for groups of related species ("clades") (e.g., clades of toxic or non-toxic *Chrysochromulina/Prymnesium* species), genus-specific probes (e.g., for *Phaeocystis* species) down to species- or even strain-level probes (e.g., *Chrysochromulina polylepsis*). These hierarchical probes make it easier to analyse field samples using an approach where first higher level probes can be applied to the samples and then, depending on these results, only probes of a corresponding lower level can be used, therefore, reducing the number of necessary experiments. For example, if the first round of testing showed an unknown species to be a dinoflagellate, then all available lower probes for other groups don't have to be tested.

Second, the use of probes for rDNA allows them also to bind to the rRNA of ribosomes in-situ, making it possible to use fluorochrome-labelled probes in whole-cell hybridisation experiments (FISH). The thousands of ribosomes provide enough targets for probe binding and therefore, strong enough signals to be detected. Using DNA as the target region, small oligonucleotide probes like these are normally not sufficient to be detected by in-situ hybridisation, even if some experiments showed otherwise. Ribosome-bound probes can then be detected by epifluorescence microscopy or automated detection systems like flow cytometers, giving the scientist the clear identification of the target cell(s) even in mixed populations together with all other information these instruments can deliver, like cell size, shape autofluorescence etc..

5.2. Probe Design

The design of molecular probes is a process that depends heavily on modern sequence databases and computer analysis programs. An alignment of e.g., rDNA sequences, consisting of the target species' sequence and at least several hundred non-target sequences is analysed by a computer program to find signature sequences for probe design. A program especially designed for this task is called ARB, which was developed by the group of W. Ludwig at the TU München, Germany. It searches for all possible unique regions using the parameters chosen by the user (probe length, GC content, etc.) and gives also lists of species inside the alignment with the fewest number of mismatches at the probe position to the target species. In some cases a single mismatch is enough to prevent probe binding to non-target species, but of course more mismatches are preferred to get a specific probe. By weighting the possible probes, the ARB program helps the user to choose one for further testing.

The next probe tests "in silico" involve the comparison of its sequence to standard databases like GENBANK, EMBL or RDP II to look for homologies to species not present in the original alignment and the analysis of the probe for possible internal loops or self-annealing that would prevent the probe to bind to its target sequence. In case of the planned use of the probe in *in-situ* experiments it is also helpful to check the probe's position on the rRNA secondary structure. It has been found that some regions of the 16S rRNA molecule of *E. coli* quenched the signal strength of fluorochrome-labelled probes to different amounts. The reason for this might be the different accessibility of the molecules regions due to stems and loops or because certain areas are covered by ribosomal proteins. Therefore, if it is possible to choose between probes for a target, one should go for those binding to already successfully tested regions. Other experiments also showed that all probes must be empirically tested before discarding because site accessibility is not strictly comparable between all species.

The following practical tests involve hybridizations of the probe to the target species and to non-target species that are taxonomically related and to those who showed homologies in the probe binding region, which is not necessarily the same, to establish the most stringent conditions for the probe to be specific.

5.3. Detection Methods

Probes could be used with different labels and different detection systems making it possible to address different kinds of questions. All kind of probes, even those not of rDNA origin could be used in DNA dot blot assays, where whole DNA or PCR products of target genes were spotted onto membranes and hybridized to chemiluminescent or radioactive-labelled probes. The advantage

of this technique is the possibility to spot a larger number of samples onto one membrane – an advantage that is even more true for DNA chips which are addressed later – and to strip and re-use it a few times with different probes so that plenty of samples could be analysed in a short time. The disadvantages on the other hand are the time-consuming process of DNA extraction and sometimes amplification by PCR to get a detectable signal strength, e.g., for field samples with low organism numbers, as well as the longer hybridisation times compared to in-situ methods. In addition, the quantification of non-radioactively labelled probe signals is still not satisfying, suffering from problems like low sensitivity of most detectors, non-linear relationship between amount of sample and signal strength because of the enzymatic step of the reaction and possible unequal amplification of different DNAs in a sample by PCR. But even with these drawbacks, the use of probes with this kind of detection method is a good way to address certain questions and has its examples in phytoplankton research as works on the toxic dinoflagellate *Alexandrium tamarense* shows.

When extracted DNA is available, another method for the use of oligonucleotide probes is as PCR primers. A specific oligonucleotide in combination with a matching primer from a highly conserved region of the same gene should only amplify a product if the DNA comes from the species the oligonucleotide probe was designed for. But it should be taken into account, that the requirements for a working PCR primer differ from those of a hybridisation probe and one to two mismatches may not be enough for a specific amplification, and the mismatches should be shifted to the 3' end of the oligonucleotide to prevent PCR amplification in non-target DNA. Nevertheless, when a probe can be used this way, the method is much faster than a dot blot hybridisation in detecting the presence of a certain type of organism, but quantification is even more difficult with the additional requirement of quantitative PCR.

The most routine method of detection is by whole-cell hybridisation (*in-situ* hybridisation) with fluorochrome-labelled probes. For this, cells are fixed by different reagents (formalin, paraformaldehyde, ethanol, etc.) to stabilise it and make it permeable for the probe. The hybridisation itself is a fast process of one to three hours with no mayor differences to other "common" hybridizations (salt-containing hybridisation buffer with detergents, hybridisation at a specific temperature, followed by a few washing steps). Stringency is also laid down by known parameters, like hybridisation and washing temperature, salt concentration and the buffer's formamide concentration). Afterwards, the samples can be analysed by different methods like epifluorescence microscopy or flow cytometry. Whereas this method normally lacks the advantage of processing numerous samples at the same time, the fact that the cells stay more or less intact gives an additional feature for identification. Viewed under the light microscope, cell shape and other morphological features can help to identify the questionable organism (Figure 3).

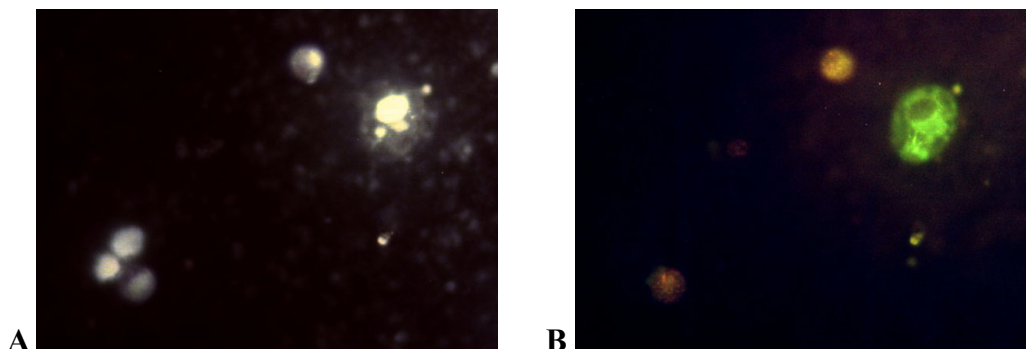


Figure 3: *In situ* hybridization of a water sample from the Irish Sea with a molecular probe specific for dinoflagellates. The staining on figure 3a shows all cells, while figure 3b shows only one cell with a positive green signal identifying it as a dinoflagellate. The other cells show only weak yellow autofluorescence.

Microscopy is not the only way to detect fluorescence in cells. Another method of detection is by flow cytometry, which is not per se a molecular biological method, but can be used in combination with molecular probes to great advantage. In flow cytometry cells are driven under pressure through a small nozzle in a way that the sample is surrounded by a sheath fluid. Single cells in the liquid stream are hit by one or more laser beams and the emitted light from the cell is analysed. Parameters determined in the first place concern the size and the geometry of organisms in the beam by means of the "forward" and "side" scatter of the laser beam. Additionally, fluorescence of different wavelength emitted by the organisms in the beam can be detected and measured, including those of a hybridized fluorochrome-labelled probe. Naturally occurring autofluorescence, e.g., from photosynthetic pigments, is an additional parameter for identifying organisms.

The advantage of flow cytometry clearly is the speed of analysis (more than 10 000 cells per second can be analysed), the accuracy of the measurement and the possibility to sort individual cells. Currently attempts are under way to develop artificial neural networks that allow an automated species recognition in natural samples using flow cytometry data. A well equipped machine, however, costs in the range of hundreds of thousands of Dollars and thus is not affordable for small labs. Also it is yet not possible to perform the detection of a particular species using in-situ hybridisation with living cells. Therefore, these cells can not be sorted alive and brought into culture, but it is possible to sort a living population of interest without preceding hybridisation and then analyse a sub-set with the probes later on.

For the application of probes, e.g., to determine the algal species composition inside a special oceanic area or to find traces of specific, toxic species in the frame of an early warning system for harmful algal blooms, it is necessary to collect and analyse a large number of samples. As mentioned previously, monitoring programs for harmful algae blooms must take place world-wide and there is a need for routine analysis of this kind or another. Whole-cell hybridisation in combination with fluorescence microscopy is clearly not fast enough to serve this task, whereas an analysis by flow cytometry is. Nevertheless, even with the help of an automatic sample loading system, the flow cytometer measures only one sample after the other which needs some time for a series when a count takes five to ten minutes per sample. Therefore, techniques that look at multiple samples at the same time are much better. DNA dot blots do this, but the technique is difficult to automate because of blotting, membrane handling, and later detection of signals. Still, there are some possible methods to combine the automation of sample processing with a fast detection system for multiple samples.

In-situ hybridisation could for example be done in microtiter plates and detection of fluorescence signals by using a plate reader. Better quantification without cell lost by centrifugation steps can be received by whole-cell hybridisation on filter membranes and analysing the resulting signals automatically with a solid phase cytometer, e.g., ChemScan (Chemunex, Ltd). This is done so far mostly with bacteria but will soon be adapted for algae.

Even greater numbers of samples and/or probes can be analysed with DNA microarrays/-chips. There are different kinds of techniques available at the moment for this, with sample numbers per

chip in the range from a few dozen up to ten thousands. The chip's size is normally that of a microscopic slide or even less. To all different systems it is common that DNA, mostly oligonucleotides, are spotted and bound onto a chip, often a coated glass slide, then hybridized to a fluorescent labelled sample, washed and the bound sample detected by its fluorescence in a special reader. The advantage of this method is the huge number of samples that can be analysed in one experiment and the large scale of automation and therefore time saving by using automated chip spotter and reader. Of course, this extensive use of robotics in the lab has its price which is even higher than the one for a flow cytometer. Also, to speak in the favour of the flow cytometer, it gives more and different information about the cells in a sample than plain readers for microtiter plates or DNA chips and can therefore make it easier to identify specific species.

All these methods have, despite their high set up costs for the machines, the disadvantage of requiring quite bulky pieces of equipment, hence, making them difficult to be used in the field and on-board ship. For this purpose, the company Inventus Biotech GmbH (Münster, Germany) is developing a handheld device for the quick and easy detection of toxic algae. In this case one or more specific probes are bound onto a disposable carbon chip and the presence and even quantity of the algae can be determined by hybridisation of a crude RNA extract to it and the resulting changes in the conductivity of the chip.

As it can be realised, there are numerous techniques possible for analysing multiple samples with specific probes automatically and more are surely to come. With them the way is open for mass screening of water samples for the detection of interesting marine species like toxic algae, even as there are still some problems to be solved and methods to optimise before they can be routinely used for this kind of purpose.

5. Conclusions

Molecular techniques can enhance our understanding of phytoplankton biodiversity in an environment as vast as the world's oceans and in organisms so tiny that they can only be reliably counted using flow cytometry. Phylogenetic diversity can be recovered without dependence on more traditional, often biased, preservation or culturing methods. Molecular techniques can reconstruct the phylogenetic history of a group and can document the spatial and temporal structuring of genetic diversity, i. e., biodiversity below the species level. A variety of molecular tools may need to be invoked in order to find the resolution needed to separate species, populations or individuals. The incorporation of all facets of the biology of the phytoplankton is essential to formulate a multidisciplinary definition of a species and to reconstruct its phylogenetic history. The potential for recognising genetic individuality at the DNA sequence level is only just being realised and its use in clustering individuals into biologically meaningful groups reflecting their overall relatedness will provide new avenues for understanding the role that phytoplankton play in structuring the marine ecosystem in both time and space.

Acknowledgements

This work was supported in part by grants from the Natural Environmental Research Council (GR3/8139 and GR9/7159), the European Commission (EHUX- MAST-PL92-0058, AIMS MAS3-

CT97-0080, PICODIV EVK3-CT-1999-00021), and the German government BMBF (TEPS - 03F0161A) to LKM.

Bibliography & Internet Resources

Amann R.I. (1995). In situ identification of micro-organisms by whole cell hybridization with rRNA-targeted nucleic acid probes. In: *Molecular Microbial Ecology Manual* **3.3.6.**, pp 1-15. Dordrecht, NL: Kluwer Academic Publishers. [Detailed protocols for development and application of molecular probes.]

Doyle R.W. (1975). Upwelling, clone selection and the characteristic shape of nutrient uptake curves. *Limnology and Oceanography* **20**: 487-489. [First to report that plankton populations must consist of multiple overlapping genotypes.]

Department of Genetics, University of Washington, USA.

[Http://evolution.genetics.washington.edu/phylip/software.html](http://evolution.genetics.washington.edu/phylip/software.html)

[A list of nearly 200 different software packages for phylogenetic analyses with comments and links for downloads. All programs mentioned in this chapter are listed here.]

Gillet E.M. (1999). *Which DNA Marker for Which Purpose? Final Compendium of the Research Project Development, optimisation and validation of molecular tools for assessment of biodiversity in forest trees in the European Union DGXII Biotechnology FW IV Research Programme Molecular Tools for Biodiversity.* [Http://webdoc.sub.gwdg.de/ebook/y/1999/whichmarker/index.htm](http://webdoc.sub.gwdg.de/ebook/y/1999/whichmarker/index.htm)

[Good overview over different types of molecular markers and their use. Concentrates on forest genetics, but a lot of information is also of interest for a broader scientific community.]

Hillis D.M., Moritz C., Mable B.K. (1996). *Molecular Systematics*, 655 pp. Sunderland, MA, USA: Sinauer Associates, Inc. [Detailed documentation of the theoretical basis for the many phylogenetic packages available as well as detailed explanations of widely used molecular methods]

Karp A., Isaac P.G., Ingram D.S. (1998). *Molecular Tools for Studying Biodiversity*, 498 pp. London, UK: Chapman & Hall. [Excellent compilation of all kinds of molecular methods usable for studying biodiversity. Includes step-by-step laboratory manuals.]

Muyzer G. (1999). DGGE/TGGE: a method for identifying genes from natural ecosystems. *Current Opinions in Microbiology* **2**:317-22. [This review describes the methods of DGGE and TGGE in detail.]

National Center for Biotechnology Information, National Institute of Health, USA.

[Http://www.ncbi.nlm.nih.gov/](http://www.ncbi.nlm.nih.gov/)

[This server offers free use of sequence comparisons by BLAST and access to databases like GenBank.]

Ormond R.F.G., Gage J.D., Angel M.V. (1997). *Marine biodiversity: Patterns and processes*, 460 pp. New York, NY, USA: Cambridge University Press.

[Overview of the range of biodiversity that has been reported in wide variety of marine habitats from the plankton to the deep-sea benthos]

Parker P.G., Snow A.A., Schug M.D., Booton G.C., Fuerst P.A. (1998). What molecules can tell us about populations: choosing and using a molecular marker. *Ecology* **79**: 361-382. [Overview and detailed comparison of the most common molecular marker types.]

Ribosomal Database Project II, Center for Microbial Ecology, Michigan State University, USA.

[Http://www.cme.msu.edu/rdp/html/index.html](http://www.cme.msu.edu/rdp/html/index.html)

[This website provides ribosome related data services, including online data analysis, rRNA derived phylogenetic trees, and aligned and annotated rRNA sequences.]

Schwieger F., Tebbe C.C. (1998). A new approach to utilize PCR-single-strand-conformation polymorphism for 16S rRNA gene-based microbial community analysis. *Applied and Environmental Microbiology* **64**:4870-4876. [This publication details the SSCP technique.]