Minireview

Metagenomics: Access to Unculturable Microbes in the Environment

Nobutada Kimura

1 Institute for Biological Resources and Functions, National Institute of Advanced Industrial Science and Technology (AIST), 1-1-1, Higashi, Tsukuba, Ibaraki 305-8566, Japan

(Received July 12, 2006—Accepted October 12, 2006)

More than 99% of the microbes that exist in the environment cannot be cultivated easily. Thus, most of the microbes in the environment have not been described and accessed for biotechnology or basic research. "Metagenomics," which is the culture-independent genomic analysis of the microbial community, attempts to overcome these difficulties. This technology has the potential for providing insight into the functional dimensions of environmental genomic datasets and will help to achieve a major goal of environmental microbiology, the complexities of microbial community function and interaction among these microbes. This article gives an overview of the current knowledge of metagenomic technologies and presents examples of studies that have used these techniques. Furthermore, future technologies related to understanding the metagenome are discussed.

Key words: metagenomics, unculturable microbes, microbial community

Introduction

Current research indicates that more than 99% of microbes in the environment cannot be easily cultivated. Thus, most microbes have not been described and accessed for biotechnology or basic research. Several DNA-based technologies have been developed to overcome these difficulties. However, only part of a gene can be amplified by PCR-based methods, and the sequence data provide few details about genetic information and any functional role in the microbial community. Furthermore, PCR amplification requires that sequence information be available prior to the design for the primers. To overcome these limitations, a new technology, Metagenomics, has been under development since the late 1990's. Metagenomics is the culture-independent genomic analysis of microbial communities. The technology was named after the term used by Handelsman to explain the statistical concept of meta-analysis and genomics. Metagenomic analyses are usually initiated by the isolation of DNA from environmental samples (Fig. 1). To increase the efficiency of gene screening in metagenomic samples, many laboratories isolate metagenomic DNA from a microbial community after enrichment in the laboratory. DNA isolation and purification is followed by the construction of DNA libraries using suitable cloning vectors, such as plasmid, cosmid, fosmid and bacterial artificial chromosome (BAC) vectors. In most research, Escherichia coli has been used as a host strain. Construction of the metagenomic library is followed by screening for novel genes or large-scale sequencing projects. Two major strategies have been used to identify novel biocatalysts or genes related to the production of chemical compounds in metagenomic libraries. Function-based screening for novel genes in metagenomes has been performed by detecting of the enzymatic activity on the plates. Recently, functional screening has been performed with a high-throughput screening technology using picking and pipetting robotics. Sequence-based screening has been performed by the PCR method with oligo primers targeting
Fig. 1. Construction and analysis of metagenomic libraries from environmental samples. The metagenomics involved constructing a DNA library and analyzing the functions and sequences in the library.
the conserved region of gene families which were designed based on the sequence information in databases. Furthermore, the complete and random sequencing of metagenomic libraries has presented useful information about novel genes. In addition, large-scale sequencing projects have provided genomic information on unculturable microbes in the environment. In this review, I describe methodological advancements and give examples of studies that have used these techniques. Furthermore, I discuss what future technology these studies could explore.

**Genome and gene enrichment**

Researchers have been able to isolate novel genes or operons from mixed laboratory cultures or enrichments much more frequently in some metagenome studies than in others. To increase the portion of desired clones in a library, several strategies have been designed to concentrate the genes of interest before cloning. Nucleic acid-based techniques have been developed that exploit the physical properties of stable isotopes to enrich a library with certain genomes. Stable-isotope probing (SIP), which involves providing a C13-labeled substrate to soil microbes, has been successfully used for labeling DNA and RNA. Density gradient centrifugation can then be used to separate the labeled nucleic acids. Furthermore, low-biomass samples from the environment yield DNA amounts that have limited use for direct, native analyses and screening. To solve these problems, genomic enrichment strategies have been developed to amplify a relatively low portion of the total DNA. The multiple displacement amplification (MDA) method was used to amplify whole genomes from environmental, contaminated, and subsurface sediment samples with ø29 DNA polymerase which is widely used for rolling-circle amplification of plasmids and circular DNA templates and is extremely sensitive, having been shown to amplify DNA of up to 70 kb. As a result of 16S rRNA-based community, small-subunit rRNA analysis revealed a relatively even distribution of species across several major phyla after amplification.

Phage-display expression libraries provide a cloning and expression system allowing the display of functional cDNAs or other gene products on the surface of filamentous phages. This method is based on the affinity of the surface-displayed protein for an immobilized ligand. Although this method limits the expression of proteins, it has the potential to make rare DNA sequences more abundant in the metagenome.

**Assessing gene diversity**

To select suitable environmental samples for metagenomic analyses, molecular techniques have been used to assess gene diversity in potential samples. A microarray was constructed to determine the potential of DNA array technology for assessing functional gene diversity and distribution. This gene array reveals differences in the apparent distribution of nir, amoA, and pmoA gene families in sediment and soil samples, indicating that microarray hybridization has the potential to reveal the functional gene composition of natural microbial communities. Microarray technology could also be used for the pre-selection of genes in metagenomic libraries.

Suppressive subtractive hybridization (SSH) has been applied as a comparative method to study microbial diversity and functional differences in the genomic content of two different environmental communities. A pilot-scale analysis of metagenomic DNA from communities of microorganisms in the rumen revealed that SSH is an effective method of identifying unique genes present in one complex microbial community and absent from another, and to isolate unique DNA fragments present in rumen metagenomic samples.

"Metagenomic profiling" offers an effective approach to rapidly characterizing many clones and identifying the clones corresponding to unidentified species of microorganisms. This method involves hybridization of the library with genomic DNA of various reference strains and bacterial isolates from the community. In addition, DNA derived from as-yet-uncultivated organisms can be identified by hybridization with metagenomic DNA. Furthermore, differential expression analysis (DEA), which targets transcriptional differences in gene expression, is considered to be a future technology for the detection of target genes in environmental samples.

**Extraction of DNA for the metagenomic library**

Two parameters have to be carefully considered when constructing metagenomic libraries (Fig. 1). The first consideration is that the large size of the metagenome requires improved cloning efficiency, so that the clones in the gene library will provide an acceptable representation of the entire metagenome. The second consideration is the size and organization of gene cluster. A typical sized of gene cluster requires the cloning of relatively large DNA fragments (in excess of 100 kb) in order to improve the probability of isolating an entire operon. To resolve these prob-
lems, a method for isolating high molecular weight DNA fragments from soil has been developed\(^{85,125}\). Protocols have been designed for isolating DNA from soil and sediment samples, which can be classified as direct and indirect extraction procedures\(^{88,80}\). Soil DNA extraction techniques used to be based on the in situ lysis of bacteria in soil prior to the recovery and purification of DNA (direct DNA isolation). In order to protect DNA from physical forces that otherwise result in DNA shearing, techniques for isolating DNA from cells that are first separated from the soil or sediment matrix before lysis (indirect DNA isolation) have been developed and allow much larger fragments to be obtained, because the cells can be immobilized in low melting point (LMP) agarose and lysed in situ. This indirect DNA isolation method was successfully used to purify large DNA fragments from oceanic bacterioplankton\(^ {113}\). Esther and coworkers compared the two DNA isolation strategies (direct DNA isolation and indirect DNA isolation) with respect to their suitability for constructing a metagenomic library, and to examine the yield, molecular mass, and cloning efficiency of environmental DNA using these strategies\(^ {39}\). Although indirect isolation methods yield 10–100 fold less DNA than direct isolation methods, the bacterial diversity of DNA recovered by the indirect DNA isolation method was distinctly higher. Furthermore, a combination of the indirect extraction of cells and a nylon capture extraction technique, which is an established method for physically separating bacterial cells from soils and sediments by means of their buoyant density, allowed fragments in excess of 1 Mb to be purified from soil followed by lysis of a biomass immobilized in agarose plugs\(^ {39}\).

**Library construction**

The classical approach includes the construction of small insert libraries (<10 kb) in a standard sequencing vector and in *Escherichia coli* as a host strain (Fig. 1). However, small insert libraries do not allow the detection of large gene clusters or operons. Furthermore, it has been estimated that more than 10\(^ 7\) clones (5 kb inserts) would be required in order to represent the collective genomes, i.e., the metagenome, of the several thousand different species that are typically present in a soil sample, assuming the idealized case of all species being equally abundant\(^ {49}\).

To overcome this limitation, large insert libraries, such as cosmid DNA libraries with an insert size of approx. 40 kb, and fosmid DNA libraries with an insert size of approx. 40 kb, have been employed to develop a metagenomic library. The bacterial artificial chromosome (BAC) vector has also been considered a powerful tool for the construction of large insert libraries (<200 kb) even since Rondon and coworkers reported its application to prokaryotic biology as an approach for studying gene function\(^ {90}\). Several researchers have succeeded in using a BAC vector to construct libraries of genomic DNA isolated directly from environmental samples\(^ {28,96,03,111}\).

Host-vector systems, composed of *Escherichia coli* strains and vectors, have been used to isolate and clone environmental DNA and screen for novel genes. Although interesting and novel activities have been expressed and identified in this host, the range of bacterial hosts needs to be expanded in order to capture additional expression capability. Other strains, such as *Streptomyces lividans*, *Rhizobium leguminosarum*, and *Pseudomonas putida*, with different expression capabilities have been employed to identify genes involved in the biosynthesis of novel antibiotics and the *trp* operon\(^ {28}\). In addition, new BAC shuttle vectors for transferring large fragments of environmental DNA from *E. coli* to both *S. lividans* and *P. putida* by high-throughput conjugation have been developed. By using this shuttle vector, Martinez found that genes involved in making different antibiotics were expressed different in *E. coli*, *P. putida* and *S. lividans*\(^ {70}\).

A mutant bacterium would be an appropriate recipient for the cloning of functional genes. Li screened individual cosmids from a metagenomic library on the basis of their ability to correct defined Trp mutants of *R. leguminosarum* and/or of *E. coli*\(^ {23}\). In this experiment, a library of ~120,000 clones containing DNA obtained directly from bacteria in a wastewater treatment plant (WWTP) were cloned in the wide host range cosmid pLAFR3.

**Approaches to metagenomic analysis**

(1) **Sequence-based screening**

Metagenomic libraries can be screened with phylogenetically relevant markers in order to link the genomic fragment to specific taxa. PCR amplification of a phylogenetic anchor, such as the 16S rRNA gene and archaeal DNA repair gene radA, is being used to obtain information about the organisms from which these clones were derived (Table 1). The potential of this strategy was publicized when Beja isolated a novel variant of a light-driven proton pump, the proteorhodopsin, that was identified on a large environmental clone linked to a 16S rRNA gene of an uncultivated marine gamma-proteobacterium\(^ {41}\). Furthermore, Stein reported that a fosmid clone which contained an archaeal small subunit ribosomal RNA gene was identified from a
fossid DNA library, which was prepared from a marine picoplankton assemblage collected in the eastern North Pacific[14].

PCR has been applied to the isolation and detection of novel genes in environmental samples. However, PCR has a potential problem when it comes to unknown genes residing in uncultured organisms, because the isolation of these genes is limited by a lack of sequence information. Furthermore, only partial genes are recovered when PCR is directed at conserved sequence motifs. To overcome these problems, Stokes reported a novel strategy for recovering complete open reading frames from environmental DNA samples. PCR assays were designed to target the 59-base element family of recombination sites that flank gene cassettes associated with integrons[13]. Using such assays, diverse gene cassettes were amplified from the environmental DNA samples tested, and these gene cassettes contained complete open reading frames, the majority of which were associated with ribosome binding sites.

(2) Function-based screening

Functional screening of metagenomic libraries have identified both novel and previously described enzymes (Table 2). A function-driven analysis is initiated by the identification of clones that have a desired trait, follow by the characterization of the isolated clones by sequencing and biochemical analysis. Although the classical isolation of active clones was performed with an in vitro system, Williamson designed a high-throughput “intracellular” screen to identify clones of interest, designated METREX, in which metagenomic DNA is in a host cell containing a biosensor for compounds that induce bacterial quorum sensing[20,30]. If the metagenomic clone produces a quorum-sensing inducer, the cell produces green fluorescent protein (GFP) and can be identified by fluorescence microscopy or captured by fluorescence-activated cell sorting. Kimura and coworkers demonstrated that metagenomic clones producing quorum-sensing inducers were isolated from soil metagenomic libraries by the METREX method (Kimura, N., B.R. Borlee, H.K. Allen and J. Handelsman, in preparation). Another “intracellular” screening method is SIGEX (substrate-induced gene expression screening), which is based on the knowledge that catabolic-gene expression is generally induced by relevant substrates and, in many cases, controlled by regulatory elements situated proximate to catabolic genes[48,124]. An operon-trap gfp-expression vector available for shotgun cloning that allows for the selection of positive clones in liquid cultures by fluorescence-activated cell sorting has been constructed and the utility of SIGEX was demonstrated by the cloning of aromatic hydrocarbon-induced genes from a groundwater metagenomic library.

Differential display (DD) is an alternate technique that can be used for the discovery of bacterial genes, requiring neither a genetic selection or screen nor the presence of highly conserved genes. The DD technique is used to compare the mRNA pools from cells grown under different

<table>
<thead>
<tr>
<th>Source</th>
<th>Number of 16S rRNA genes</th>
<th>Number of clones</th>
<th>Insert size (kb)</th>
<th>Total DNA (G bp)</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bacteria</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Marine</td>
<td>38</td>
<td>32,000</td>
<td>10-20</td>
<td>0.48</td>
<td>109</td>
</tr>
<tr>
<td>Polychaete symbionts</td>
<td>2</td>
<td>n.r.</td>
<td>35.2, 38</td>
<td>n.r.</td>
<td>21</td>
</tr>
<tr>
<td>Marine</td>
<td>2</td>
<td>3,200</td>
<td>39.5</td>
<td>n.r.</td>
<td>81</td>
</tr>
<tr>
<td>Soil</td>
<td>28</td>
<td>24,400</td>
<td>42</td>
<td>1.2</td>
<td>73</td>
</tr>
<tr>
<td>Soil</td>
<td>1</td>
<td>25,278</td>
<td>35-40</td>
<td>0.95</td>
<td>90</td>
</tr>
<tr>
<td>Soil</td>
<td>6</td>
<td>56,000</td>
<td>33-44</td>
<td>2.2</td>
<td>91</td>
</tr>
<tr>
<td>Soil</td>
<td>4</td>
<td>3552</td>
<td>40</td>
<td>2.2</td>
<td>127</td>
</tr>
<tr>
<td>Spring sediment</td>
<td>35</td>
<td>19,200</td>
<td>35</td>
<td>n.r.</td>
<td>33</td>
</tr>
<tr>
<td>Spring sediment</td>
<td>2</td>
<td>41-49</td>
<td>n.r.</td>
<td>83</td>
<td></td>
</tr>
<tr>
<td>Sediment</td>
<td>2</td>
<td>10,000</td>
<td>40</td>
<td>n.r.</td>
<td>82</td>
</tr>
<tr>
<td>Archaea</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Marine</td>
<td>n.r.</td>
<td>6,240</td>
<td>80</td>
<td>n.r.</td>
<td>4</td>
</tr>
<tr>
<td>Marine</td>
<td>5</td>
<td>7,200</td>
<td>40</td>
<td>2.9</td>
<td>3</td>
</tr>
<tr>
<td>Sponge symbionts</td>
<td>15</td>
<td>n.r.</td>
<td>40</td>
<td>n.r.</td>
<td>107</td>
</tr>
<tr>
<td>Marine</td>
<td>4</td>
<td>3552</td>
<td>40</td>
<td>0.14</td>
<td>114</td>
</tr>
<tr>
<td>Soil</td>
<td>100</td>
<td>7567.8</td>
<td>33-45</td>
<td>3.02</td>
<td>120</td>
</tr>
</tbody>
</table>

n.r.: not reported

Table 1. Metagenomic analysis based on functional screening 16S rRNA sequences
physiological conditions and involves the reproducible amplification of DNA fragments from the mRNA population at arbitrary sites by reverse transcription (RT) followed by PCR (RT-PCR). This DD approach did not use to be applied to prokaryotes, which lack stable poly (A) tails, but is applied now. Brzostowicz has shown that a high-throughput approach to DD, using a large set of arbitrary oligonucleotides to initiate RT-PCR, resulted in the identification of cyclohexanone oxidation genes from prokaryotes in a mixed microbial community derived from a wastewater bioreactor^{20}.

**Metagenomics for community analysis**

Metagenomic libraries of environmental DNA enable the exploration of the phylogenetic and metabolic diversity of microbes in the environment without cultivation or PCR (Table 1)^{24,121}. Stein and coworkers have demonstrated the utility of capturing community DNA in the form of BAC libraries, and linking phylogenetic and functional information...
tion within specific BAC clones\(^{45}\). Beja and coworkers have discovered a 140-kb BAC clone from a marine picoplankton genomic DNA library that contained both an rRNA operon (SAR86 clade) and an open reading frame (ORF) with high homology to bacteriorhodopsin\(^{3,4,7}\). This strategy has also recently been used by Quaiser to identify a fosmid clone derived from an uncultured soil crenarchaeote, revealing physiologically relevant functional genes adjacent to an rRNA operon\(^{40}\).

It is also important to compare the phylogenetic member of metagenomic libraries to other measures of community structure, such as 16S rRNA clone libraries and fluorescence in situ hybridization (FISH), in order to determine if any of the dominant groups of bacteria are missing. In addition, the phylogenetic information in metagenomic libraries provides another view of community structure without the biases of PCR-dependent approaches\(^{29}\). Liles and coworkers demonstrated the results of a census of rRNA genes within a constructed BAC library to characterize potential biases involved in the construction of metagenomic libraries (e.g., cell lysis, restriction digestion, and cloning)\(^{13}\).

The development of sequencing technology allows us to sequence entire metagenomes of environmental samples\(^{3,10,18,122,123,126}\). Using the random shotgun sequencing of DNA from a natural acidophilic biofilm which was dominated by a small number of species populations, reconstruction of near-complete genomes of *Leptospirillum* group II and *Ferroplasma* type II, and a partial recovery of three other genomes, have been reported\(^{27}\). Analysis of the gene complement for each organism revealed the pathways for carbon and nitrogen fixation and energy generation, and provided insights into survival strategies in an extreme environment. Furthermore, Venter and coworkers conducted a massive sequencing project focused on the marine microbial community of the Sargasso Sea\(^{79,126}\). A total of 1.045 billion base pairs of nonredundant sequence was generated, annotated, and analyzed to elucidate the gene content, diversity, and relative abundance of the organisms within these environmental samples. These data are estimated to derive from at least 1,800 genomic species based on sequence relatedness, including 148 previously unknown bacterial phylotypes. Over 1.2 million previously unknown genes represented in these samples have been identified, including more than 782 new rhodopsin-like photoreceptors\(^{42}\).

Metagenomic analysis also revealed the viral diversity of the human and marine environments\(^{102}\). A genomic analysis of two uncultured marine viral communities was reported by Breitbart, who found that over 65% of the sequences were not significantly similar to previously reported sequences\(^{101}\). The most common significant hits among the known sequences were to viruses. Several independent mathematical models based on the observed number of contigs predicted that the most abundant viral genome comprised 2–3% of the total population in both communities, which was estimated to contain between 374 and 7,114 viral types. Bacteriophages likely exert a strong influence on the diversity and population structure of bacterial communities in the human gut. Sequencing of the viral community from human feces also indicated that the recognizable viruses were mostly siphophages, and that the community contained an estimated 1,200 viral genotypes\(^{18}\). Phage diversity also has been revealed by metagenomic technology\(^{102}\).

**Target of metagenomic analysis**

1. **Novel functional gene and enzyme**

Metagenomics has identified a number of novel genes encoding promising biocatalysts such as lipases, esterases, chitinases, amidases, amylases, genes encoding enzymes for the metabolism of 4-hydroxybutyrase, and genes encoding the components of the biotin synthetic pathway (Table 3). Some of these enzymes have already been commercialized by venture companies such as Diversa Corporation (http://www.diversa.com). Metagenomic libraries have also been used for isolating antibiotic resistance genes from environmental samples. Riesenfeld identified nine clones expressing resistance to aminoglycoside antibiotics and one expressing resistance to tetracycline in soil metagenomic libraries\(^{46}\). Recently, Diaz-Torres and colleagues constructed metagenomic libraries from the human oral cavity and identified novel tetracycline, amoxycillin, and gentamicin resistance genes\(^{41}\). These results indicate that the bacterial community is a reservoir of antibiotic resistance genes with greater genetic diversity than previously accounted for, and that the diversity can be surveyed by a culture-independent method.

Bacteria capable of xenobiotic degradation are widely distributed in the environment\(^{62-65}\). These bacteria have evolved to utilize a variety of compounds that are present in the environment\(^{53,59,60,64,117,146}\). Hence, environmental samples are considered to be a reservoir of useful enzymes for industrial catalysis and bioremediation\(^{51}\). The metagenomic approach has also demonstrated that novel metabolic genes play an important role in the biodegradation of compounds\(^{49}\). A microbial mat from the Black Sea was analyzed by a metagenomic approach, and all genes required for the complete anaerobic degradation of benzoate (cata-
bolic island) were obtained from the mat, providing the first evidence of the capacity for such degradation \(^{69}\). In addition, metagenomics have identified a number of novel genes encoding enzymes of biodegradation, such as benzoate, the pesticide-degrading gene (\textit{linA}) product, and catechol 2,3-dioxygenase (\textit{nahH}) \(^{69}\).

In general, searches in metagenomic-driven DNA libraries have mainly focused on a rather small group of enzymes. Owing to the development of a method for the construction of large insert DNA libraries, gene clusters for catabolic enzymes have been identified and sequences of the flanking region of some clones revealed a gene or a group of genes that can be used to infer the phylogenetic affiliation of the organism from which the DNA in the clone was isolated.

(2) Natural products

A number of novel genes encoding molecules with potential for use in pharmaceutical products have been identified by metagenomics (Tables 3 and 4). The most popular genes that have been isolated with this research have been type I and type II polyketide synthases (PKSs), which are a key enzyme for synthesizing polyketide antibiotics and part of large biosynthetic gene clusters \(^{26,76,88}\). Seow and coworkers have provided the first example of the isolation of type II polyketide synthase gene homologs from an environmental community of microorganisms \(^{112}\). Two CLF (KS beta) genes were cloned from soil DNA with PCR primers derived from conserved regions of known ketosynthase (KS alpha) and ACP genes specifying the formation of polyketides. Interestingly, Piel has reported a putative antitumour pederin PKS from an uncultured bacterial symbiont of \textit{Paederus fuscipes} beetles \(^{77}\). Antitumor polyketide biosynthesis by an uncultivated bacterial symbiont of the marine sponge \textit{Theonella swinhoei} has been also reported \(^{99}\).

Metagenomics has identified a number of genes involved in the synthesis of novel antibiotics. Brady reported that a clone found in a soil metagenomic library produces deoxyviolacein and the broad spectrum antibiotic violacein \(^{59}\). Turbomycin A and B, which are two colored triaryl cation antibiotics, and their synthesizing genes have been isolated from a soil metagenomic library \(^{101}\). In addition, a number of genes involved in producing other natural products, such as isocyanide, indirubin, indigo, antimicrobial small molecules, long-chain \textit{N}-acyltyrosine synthases, and long-chain \textit{N}-acyltyrosine synthases have been isolated and characterized by the metagenomic approach (Table 3).

A number of natural products are produced by Gram-positive bacteria, such as \textit{Streptomyces} species. Although \textit{E. coli} is the preferred host for the cloning and expression of any metagenome-derived genes, other hosts have been employed to identify genes involved in the biosynthesis of novel antibiotics. Wang reported that soil metagenomic libraries were constructed by cloning large fragments of DNA isolated from soil into a \textit{Streptomyces lividans} host, and five novel compounds, terragine A, B, C, D and E, were isolated \(^{122}\). These results indicated that the cloning of microbial DNA fragments isolated from soil samples and their expression in \textit{Streptomyces} hosts can extend the spectrum of potentially useful compounds made by the host strain. As many studies on biosynthetic pathways of known natural products have shown, secondary metabolite biosynthetic pathways are tightly clustered on bacterial chromosomes, and as a result the cloning and heterologous expression of secondary metabolites from single continuous fragments of genomic DNA have been possible. Hence, large insert DNA libraries, such as BAC libraries, are preferable for the isolation of natural products.

**Recently developed technology for metagenomic analysis**

The application of metagenomic technologies to construct cDNA libraries from RNA isolated from environmental samples has been reported \(^{40}\). Owing to the presence of intron sequences, metagenomic cDNA libraries were not suitable for detecting eukaryotic genes. To overcome this limitation, the libraries were made using polyadenylated RNA isolated from total RNA from activated sludge. Sequence analysis revealed that many clones in these libraries had significant similarity to eukaryote mRNA-encoded protein sequences although open reading frames (ORFs) of up to 378 amino acids in size were identified in this study, since the process of RT-PCR amplification limits the size of inserts.

With the availability of metagenomic sequences and the increasing number of complete individual genome sequences, it is possible to apply postgenomic techniques, particularly proteomics, to complex microbial communities \(^{57,138}\). To successfully extract and purify the entire proteome from environmental samples, two-dimensional polyacrylamide gel electrophoresis and the mapping of this metaproteome were performed. Highly expressed protein spots were excised and identified using quadrupole time-of-flight mass spectrometry with \textit{de novo} peptide sequencing \(^{377}\). Metaproteomics also evaluated gene expression, identified key activities, and examined the partitioning of metabolic functions in a natural acid mine drainage (AMD) \(^{93}\). Metaproteomics is a promising technology with
Table 3. Metagenomic discovery of novel genes and natural products

<table>
<thead>
<tr>
<th>Novel genes</th>
<th>References</th>
<th>Natural products</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>4-Hydroxybutyrate</td>
<td>51</td>
<td>Antibacterial molecule</td>
<td>76</td>
</tr>
<tr>
<td>Alcohol oxidoreductase</td>
<td>67</td>
<td>Antibacterial violacein</td>
<td>13</td>
</tr>
<tr>
<td>Alcohol/Aldehyde dehydrogenase</td>
<td>135</td>
<td>Antibiotic</td>
<td>11, 12, 15, 16, 30, 31, 41</td>
</tr>
<tr>
<td>Amidase</td>
<td>38</td>
<td>Antitumour polyketide</td>
<td>86, 89</td>
</tr>
<tr>
<td>Antibiotic resistant</td>
<td>30, 31, 98, 113</td>
<td>Long-chain fatty acid ester</td>
<td>14</td>
</tr>
<tr>
<td>Amylase</td>
<td>97, 141</td>
<td>Quorum sensing inducer</td>
<td>136</td>
</tr>
<tr>
<td>Nuclease</td>
<td>103</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2,5-Diketo-D-gluconic acid reductase</td>
<td>35</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hemolytic activity</td>
<td>103</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Benzoate catabolism</td>
<td>69</td>
<td></td>
<td></td>
</tr>
<tr>
<td>β-Glucanase</td>
<td>131</td>
<td></td>
<td></td>
</tr>
<tr>
<td>β-Lactamase</td>
<td>113</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Biotin biosynthesis</td>
<td>34</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Carboxyls formation</td>
<td>66</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Catechol 2,3-dioxygenase</td>
<td>84</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cellulase</td>
<td>50</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Xylosidase</td>
<td>50</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chitinase</td>
<td>23, 70</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coenzyme B(12)-dependent glycerol and diol dehydratases</td>
<td>65</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Esterase</td>
<td>36, 96</td>
<td></td>
<td></td>
</tr>
<tr>
<td>H₂MPT-linked C₁ transfer</td>
<td>55</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Indirubin and indigo production</td>
<td>74</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Isocyanate biosynthesis</td>
<td>17</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lipase</td>
<td>34, 94</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Na⁺/H⁺ antiporter</td>
<td>77</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pesticide degradation</td>
<td>10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Polyketide synthase</td>
<td>42, 106</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4. Metagenomic analysis of targeted genes or gene families

<table>
<thead>
<tr>
<th>Target gene</th>
<th>Source</th>
<th>Host strain</th>
<th>Vector</th>
<th>insert size (kb)</th>
<th>Number of clones</th>
<th>Positive clones</th>
<th>Total DNA (G bp)</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Histidine protein kinase</td>
<td>Tubeworm symbiont</td>
<td>E. coli</td>
<td>Fosmid</td>
<td>n.r.</td>
<td>1,500</td>
<td>9</td>
<td>0.06</td>
<td>54</td>
</tr>
<tr>
<td>Hydrolases</td>
<td>Marine</td>
<td>E. coli</td>
<td>Fosmid</td>
<td>n.r.</td>
<td>n.r.</td>
<td>n.r.</td>
<td>n.r.</td>
<td>25</td>
</tr>
<tr>
<td>Photosystem II</td>
<td>Marine</td>
<td>n.r.</td>
<td>n.r.</td>
<td>80</td>
<td>6,240</td>
<td>n.r.</td>
<td>n.r.</td>
<td>143</td>
</tr>
<tr>
<td>Proteorhodopsin</td>
<td>Marine</td>
<td>E. coli</td>
<td>BAC</td>
<td>n.r.</td>
<td>n.r.</td>
<td>n.r.</td>
<td>n.r.</td>
<td>29</td>
</tr>
<tr>
<td>Polyketide synthase</td>
<td>Marine</td>
<td>E. coli</td>
<td>50</td>
<td>5,000</td>
<td>n.r.</td>
<td>0.25</td>
<td>26</td>
<td></td>
</tr>
<tr>
<td>Polyketide synthase</td>
<td>Beetle symbions</td>
<td>E. coli</td>
<td>Fosmid</td>
<td>n.r.</td>
<td>n.r.</td>
<td>8</td>
<td>n.r.</td>
<td>87</td>
</tr>
<tr>
<td>Polyketide synthase</td>
<td>Beetle and sponge symbions</td>
<td>E. coli</td>
<td>Cosmid</td>
<td>n.r.</td>
<td>n.r.</td>
<td>3</td>
<td>n.r.</td>
<td>88</td>
</tr>
<tr>
<td>Polyketide synthase</td>
<td>Soil</td>
<td>E. coli</td>
<td>Cosmid</td>
<td>60,000</td>
<td>40</td>
<td>n.r.</td>
<td>42</td>
<td></td>
</tr>
</tbody>
</table>

n.r.: not reported

which to enhance the understanding of the microbial world and link microbial community composition to function.

Web sites and Databases

Several web sites and databases have been developed for analyzing metagenomic information, including
Megx.net, a set of databases and tools that handle metagenomic sequences in their environmental context. Megx.net includes (i) a geographic information system to systematically store and analyze marine genomic and metagenomic data in conjunction with contextual information; (ii) an environmental genome browser with fast search functionalities; (iii) a database with precomputed analyses for selected complete genomes; and (iv) a database and tool to classify metagenomic fragments based on oligonucleotide signatures. All resources are freely accessible at http://www.megx.net.

A central problem with the metagenome approach is that the cloned fragments often lack suitable phylogenetic marker genes, rendering the identification of clones that are likely to originate from the same genome difficult or impossible. For the purpose of phylogenetic affiliation, the analysis of intrinsic DNA-signatures like tetranucleotide frequencies that provide valuable hints on fragment affiliation have been developed. With this application in mind, the TETRA web-service and TETRA stand-alone program have been developed, both of which automate the task of comparative tetranucleotide frequency analysis. http://www.megx.net/tetra. TETRA provides a statistical analysis of tetranucleotide usage patterns in genomic fragments, either via a web-service or a stand-alone program.

Concluding remarks and prospects

Metagenomic technology is used to study prokaryotes and eukaryotes in the environment that are, as yet, unculturable, and which represent more than 99% of the organisms in some environments. A number of subjects have been dealt with using current metagenomic technology, such as the discovery of novel natural products and enzymes, genomic analyses of unculturable bacteria, and community analyses. It is expected that the number of novel genes identified using metagenomic technology will exceed the number identified through sequencing of novel genes from isolated individual microbes. Metagenomics may provide insight into the functional dimensions of environmental genomic datasets and will help to achieve a major goal of environmental microbiology: the ability to link individual microbial species to function. However, metagenomics is still a developing technology with limitations to be overcome, such as gene enrichment, DNA extraction, host-vector design, library construction, screening, and the phylogenetic affiliation of isolated genes.

Acknowledgements

This study was carried out as a part of "The Leading Research Program for International Joint Investigation," which was supported by the New Energy and Industrial Technology Development Organization (NEDO).

References

Metagenomics for Unculturable Microbes


76) MacNeil, I.A., C.L. Tiong, C. Minor, P.R. August, T.H. Gross-


