PUBLICATIONS
Identification and characterization of genes conferring salt tolerance to *Escherichia coli* from pond water metagenome

Raj Kishor Kapardar a,b, Ravi Ranjan a, Amit Grover a, Munish Puri b,1, Rakesh Sharma a,⁎

a Microbial Genomics and Biotechnology Group, Institute of Genomics and Integrative Biology, Council of Scientific and Industrial Research (CSIR), Mall Road, New Delhi 110007, India
b Fermentation and Protein Biotechnology Group, Department of Biotechnology, Punjab University, Patiala 47002, India

Abstract

Metagenomics provides culture-independent access to gene pool of the whole microbial communities. To identify genes responsible for salt tolerance in unculturable bacteria, *Escherichia coli* clones were enriched with an ability to grow at inhibitory NaCl concentrations (750 mM) from a pond water metagenomic library. From two unique clones, genes encoding for proteins with similarity to a putative general stress protein (GspM) harbouring GsiB domain and a putative enoyl-CoA hydratase (EchM) were identified to be responsible for salt tolerance. The gspM was expressed by its native promoter whereas the echM was expressed from the lacZ promoter of the plasmid. EchM was overexpressed with a hexahistidyl tag. Purified EchM showed crotonyl-CoA hydratase activity. These genes have potential application in generating salt tolerant recombinant bacteria or transgenic plants.

© 2010 Elsevier Ltd. All rights reserved.

ARTICLE INFO

Article history:
Received 1 October 2009
Received in revised form 5 January 2010
Accepted 10 January 2010
Available online 4 February 2010

Keywords:
Metagenome
Metagenomic
Salt tolerance
Enoyl-CoA hydratase
General stress protein

1. Introduction

Most bacteria possess ability to tolerate salt stress for a short period, whereas some others can grow at high salt concentration. The mechanisms responsible for salt stress tolerance and growth under high salt conditions range from induction of general stress tolerance proteins (Hillmann et al., 2006), Na⁺/H⁺ efflux pumps (Waditee et al., 2002), modification of membrane composition and fluidity (Sakamoto and Murata, 2002) to increased production and/or uptake of compatible solutes (Roberts, 2005). Various genes responsible for providing salt tolerance and ability to grow at elevated salt concentration have been identified from different bacteria (Zuleta et al., 2003) including the genes involved in synthesis and uptake of various compatible solutes (Naughton et al., 2009). These genes are important for generation of salt tolerant recombinant microbial strains and transgenic plants (Kapley et al., 1999; Klahn et al., 2009). Many bacterial genes have been successfully applied to generate salt tolerant transgenic plants after suitable modifications in their expression machinery (Chen et al., 2007; Klahn et al., 2009).

Existence of a large number of unculturable bacteria in environmental samples was realized relatively recently by microbiologists. This vast 'yet to be cultured' microbial diversity is expected to possess many unique physiological mechanisms. Many of such mechanisms have been revealed by metagenomic approaches aiming to understand physiology of these unculturable bacteria (Chauhan et al., 2009; Singh et al., 2009a). A large number of biotechnologically important genes for biocatalysts and antimicrobial compounds have been successfully identified through functional metagenomics (Sharma et al., 2005; Singh et al., 2009b). Most bacteria in their natural habitats encounter various abiotic stress conditions. Possession of distinct stress tolerance mechanisms provides them edge over their competitors to survive under these stress conditions. Pond water microbial communities withstand significant fluxes in salinity due to evaporation and dilution. Unculturable bacteria of pond water are expected to possess unique stress tolerance mechanisms. Identification of salt tolerance genes using metagenomic approach is expected to reveal new salt tolerance mechanisms of yet uncharacterized bacteria.

In an attempt to exploit the gene pool of pond water microbial community, a pond water metagenomic library was generated (Ranjan et al., 2005). Identification of two novel salt tolerance genes, encoding a putative general stress protein and an enoyl-CoA hydratase, from this pond water metagenomic library is reported here.

2. Methods

2.1. Culture conditions and metagenomic library

A list of strains and plasmids used in this study is given in Table 1. The *Escherichia coli* strains were grown in Luria–Bertani (LB)
liquid medium or on LB agar plates at 37 °C. The salt concentrations mentioned are final concentrations of salt in the medium including the salt present in normal LB medium. The media was supplemented with kanamycin (50 μg/ml) and/or ampicillin (100 μg/ml) for selection of plasmids. Metagenomic DNA isolation from pond water sample and construction of metagenomic library were described previously (Ranjan et al., 2005).

2.2. Enrichment of salt tolerant clones

One milliliter of amplified pond water library culture was inoculated in 20 ml LB broth supplemented with ampicillin (100 μg/ml) and 850 mM or 1025 mM Sodium Chloride (NaCl). The flasks were incubated at 37 °C for 24 h with shaking at 200 rpm. One milliliter of culture was then transferred to new flasks containing 20 ml of the same components. The enrichment process for the salt tolerant clones was repeated 10 times and then the cultures were plated on same medium containing agar plates. Distinct colonies appeared on the plates after 72 h incubation at 37 °C. Plasmids were isolated from ten colonies from each plate and analyzed by restriction digestion with EcoR1 and Hind111.

2.3. DNA manipulation

Plasmid isolation, restriction enzyme digestion, ligation, and competent cells preparation were carried out by standard procedures (Sambrook and Russell, 2001).

2.4. Transposon mutagenesis and mapping

Plasmid from the selected clones was used for in vitro transposon mutagenesis by template generation system (TGS, F-700) by the protocol provided by the manufacturer (Finzyme, Finland). The mutagenized plasmid was electroporated in E. coli DH10B and colonies containing transposon-mutated plasmids were selected on ampicillin and kanamycin containing LB agar plates. Randomly selected mutants were sequenced using primers against transposon ends. Complete sequence for both the strands for the inserts was obtained by sequencing the gap region by primers walking. To identify salt tolerance genes, mutants were patched with help of sterile colonies containing transposon insertion in gene conferring salt tolerance genes. Location of the transposon insertion in the selected mutated plasmids was mapped either by size of PCR product obtained by primers against transposon ends and pUC reverse (5' - TATGTTGTG-3') or pUC forward primer (5' - AGCTGGCGAAAGGGG-GATGTCG-3') or by DNA sequencing of flanking regions.

2.5. DNA sequencing and analysis

DNA sequencing was performed with big dye termination cycle sequencing kit version 3.1 (Applied Biosystems, USA) and the products were run on a ABI 3700 machine using POP6 method at The Centre for Genomic Application, New Delhi, India. Transposon ends specific primers SeqA and SeqB (supplied with the TGS kit) were used to sequence mutated plasmids. Sequence assembly and analysis was carried out using Lasergene package, version 5.07 (DNA Star, USA). Open reading frames in the assembled sequence for each clone were identified by ORF finder and the amino acid sequence of each identified ORF was used to find the closest match by BLAST without low complexity filter at NCBI website (http://www.ncbi.nlm.nih.gov). Conserved domains and patterns were analyzed using conserved protein domain database (Marchler-Bauer and Bryant, 2004) and Prosite (Hulo et al., 2004).

2.6. Subcloning

A 972 bp EcoR1-Sal1 fragment from p5B2 containing gspM gene was sub-cloned in pUC19 digested with same restriction enzymes to create pRK1. The EcoR1-Sal1 fragment was cloned in pUC18 to obtain pRK2, which contained gspM in opposite orientation to lacZ promoter. A 1277 bp long Pst1 fragment carrying echM gene was isolated from p652 and cloned into Pst1 site of pUC19, the resultant plasmids were checked for the orientation of the cloned fragment. The plasmids with orientation similar to p652 were designated as pRK3 and the plasmid with fragment in opposite orientation were designated as pRK4. In plasmid pRK3, the identified gene was located in opposite orientation to the lacZ promoter of the vector. To check if expression of gspM and echM together can result in higher salt tolerance of the recombinant E. coli strain, the Pst1 fragment from p652 containing echM gene was cloned in Pst1 site of pRK1. Orientation of the Pst1 fragment was checked by restriction digestion and the plasmid with echM gene orientation opposite to lacZ promoter was designated as pRK5 and the one in similar direction to lacZ promoter was designated as pRK6.

2.7. Overexpression and purification of EchM

The echM gene was amplified by PCR using the primers EchM for 5'-CATGCAA TGGAA CGGCTTCAAGGCCGACAAAATACG-3' and EchMexwv 5'-CCAAAGCGCTTGCCGT TTTTGACTTTGCTGCC-3' and p652 as template. Following restriction digestion with NcoI and...
6.7, culture techniques (Rani et al., 2008). These unculturable bacteria revealed that majority of bacteria cannot be cultivated by routine exploitation (Sharma et al., 2005; Dubey et al., 2006). Advances in metagenomics, ranging from better DNA isolation methods to availability of newer vectors and hosts, has improved the ability to access the gene pool of these unculturable bacteria (Desai and Madamwar, 2006; Singh et al., 2009b).

2.8. Enoyl-CoA hydratase assay

The enoyl-CoA hydratase activity was assayed by the hydration of crotonyl-CoA by slight modification of the procedure described earlier (Fukui et al., 1998). A 10 μl of diluted enzyme solution was added to 990 μl of 50 mM Tris Cl, pH-8.0 containing 250 μM crotonyl-CoA (Sigma) in quartz cuvettes with a 1.0 cm path length, and the decrease in absorbance at 263 nm was measured at 30 °C. The ε_{263} of the enoyl-thioester bond is taken to be 6.7 × 10^{3} M^{-1} cm^{-1}. Protein concentration was determined by the Bradford method with bovine serum albumin as standard.

2.9. Nucleotide sequence accession number

The sequences obtained from the clones were deposited in GenBank under accession numbers EF611421 and EF611422.

3. Results and discussion

Bacterial diversity analysis by culture-independent methods revealed that majority of bacteria cannot be cultivated by routine culture techniques (Rani et al., 2008). These unculturable bacteria represent a rich and vast genetic resource for biotechnological exploitation (Sharma et al., 2005; Dubey et al., 2006). Advances in metagenomics, ranging from better DNA isolation methods to availability of newer vectors and hosts, has improved the ability to access the gene pool of these unculturable bacteria (Desai and Madamwar, 2006; Singh et al., 2009b).

3.1. Selection of salt tolerant clones

Salt tolerant clones were enriched by successive culture of a pond water metagenomic library, containing about 532 Mb of community DNA in ~140,000 recombinants, in LB broth containing 100 μg/ml ampicillin and 850 or 1025 mM NaCl. Enriched clones were selected on agar plates containing the same medium. Three unique insert carrying plasmids were identified by restriction digestion analysis, one from 850 mM NaCl containing medium, designated as p5B2 and two plasmids from 1025 mM NaCl containing medium, designated as p6S2 and p6B4. Plasmids were re-transformed in E. coli DH10B strain and the colonies obtained were checked for growth at 850 mM NaCl containing medium. All the colonies grew in this medium indicating plasmid borne nature of salt tolerance, the control strain with pUC19 failed to grow under same conditions. Restriction analysis of p5B2, p6S2 and p6B4 revealed that they contained insert of ~2.6 kb, ~4.0 kb and ~2.3 kb, respectively.

Enrichment of the pond water library for E. coli clones with ability to grow at NaCl concentration above 750 mM resulted in isolation of three unique clones confirming that metagenomic libraries are a gene pool resource that can be exploited for identification of genes for many important functions. Fresh water ponds, lakes, and water bodies have been shown to possess unique microbial diversity including members of various unculturable groups (Humbert et al., 2009). A pond water metagenomic library was prepared earlier and novel lipolytic genes were identified by functional screening of the library, indicating incorporation of distinct and as yet uncharacterized gene pool of uncultured bacteria of pond water in this metagenomic library (Ranjan et al., 2005).

3.2. Growth characteristics of salt tolerant clones

The growth curves of the salt tolerant clones were analyzed in LB broth supplemented with ampicillin and either with no other additive or with 750 mM NaCl (Fig. 1). The clones E. coli/p5B2, E. coli/p6S2 and E. coli/p6B4 grew with different rates and achieved maximum growth in 24–28 h in medium containing 750 mM NaCl in comparison to control E. coli/pUC19 or E. coli/pLC8, which did not achieve significant growth even till 40 h (Fig. 1A). E. coli/pLC8, a recombinant plasmid carrying clone from pond water metagenomic library was taken as second negative control in all experiments, as in LB medium E. coli/pUC19 grew with slower rate in comparison to recombinant clones probably because of stress caused by β-galactosidase expression. In LB medium, the clone E. coli/p6B4 grew much slower in comparison to controls and other two clones indicating that presence of recombinant plasmid itself caused stress to this clone (Fig. 1B). The clone E. coli/p6B4 was not selected for further analysis.

3.3. Sequence analysis of the clones

The DNA sequences of both strands of inserts from p5B2 and p6S2 were obtained by sequencing transposon-flanking regions of the plasmids mutated by template generation system and primer walking. DNA sequence of 2593 bp with a G + C composition of 59.58% and 4019 bp with a G + C composition of 63.97% was obtained from p5B2. Encoded proteins from 64% to 75% were identified from p5B2. Encoded proteins, a putative response regulator and a putative ECF subfamily protein, a putative feruloyl-CoA synthase and a putative acetyl-CoA acetyltransferase with amino acid identity ranging from 64% to 75% were identified from p5B2. Encoded proteins with similarity to a putative enoyl-CoA hydratase/isomerase, a putative feruloyl-CoA synthase and a putative acetyl-CoA acetyltransferase with amino acid identity ranging between 77% and 84% were identified from p6S2 (Table 2). No nucleotide match or low amino acid identity of the encoded proteins with the known proteins in the databases indicated at possibility of the metagenomic insert in p5B2 from as yet uncharacterized bacteria. Whereas, significant nucleotide identity with Rhodotherax and Polaronomas sp. JS666 with nucleotide identity ranging from 81% to 90%.


3919
3.4. Identification of gene conferring salt tolerance from p5B2

Eight salt sensitive insertional mutants were obtained for plasmid p5B2 out of the 400 mutants screened. The clones with the mutant plasmids behaved similar to *E. coli* /pUC19 in LB broth containing ampicillin and 750 mM NaCl (data not shown). Mapping and sequencing of these negative mutants revealed that five transposons were located in an ORF encoding for 93 amino acids protein and three in the upstream region of this gene within 60 bp of the start site (Fig 2). This indicated that disruption of structural gene or putative promoter region of this ORF was responsible for the salt sensitive phenotype of the mutants. The encoded protein of about 10 kDa showed best match with a hypothetical protein of *Methylocella silvestris* BL2 with 68% amino acid identity. Many of the other matches were annotated as general stress proteins in the database. The gene was designated as *gspM* because of its similarity with genes encoding for putative general stress proteins and metagenomic origin.

*E. coli*/pRK1 showed growth similar to *E. coli/p5B2* in 750 mM NaCl containing LB medium (Supplementary data), confirming that the salt tolerance was conferred by *gspM*. To confirm if the *gspM* was expressed by its own promoter or with lacZ promoter from the vector, *gspM* gene was cloned in opposite orientation to the *lacZ* promoter in pRK2. *E. coli/pRK2* also showed similar growth to *E. coli/pRK1* in 750 mM NaCl containing LB medium (Supplementary data) indicating expression of *gspM* by its native promoter.

GspM had a complete domain of general stress protein family GsiB (COG3729). InterProScan search revealed similarity of GspM with members of Family PD027049 (ProDom release 2005.1). The most well characterized member of these families is glucose starvation induced protein (GsiB) of *Bacillus subtilis*, which is a hydrophilic protein of 123 amino acids and is composed almost entirely of five repeating motifs (Stacy and Aalen, 1998). Sequence analysis of GspM by radar repeat finder revealed that it was not formed of perfect repeats similar to GsiB but a motif R(K/T)GG was found repeated three times (Supplementary data) in GspM. Sequence alignment of GspM and its homologues with proteins from GsiB and PD027049 families did not reveal significant similarity between these proteins and GspM. Phylogenetic analysis revealed that GspM and its homologues formed a distinct clade separate from the clades formed by GsiB, YciG or Late embryogenesis abundant proteins homologues (Fig. 3). GspM grouped with homologous proteins from alpha-proteobacteria indicating its possible origin from an alpha-proteobacteria of pond water sample.

**Fig. 1.** Growth curves of *E. coli* strains, (○) *E. coli/p5B2*, (●) *E. coli/p6S2*, (♦) *E. coli/pUC19*, (●) control *E. coli/pLR8* in (A) LB medium with 750 mM NaCl and (B) LB medium. Experiments were performed in triplicates, values are averages and standard deviations are shown as error bars.

---

**Table 2**

<table>
<thead>
<tr>
<th>Clone</th>
<th>Insert size (G + C %)</th>
<th>Open reading frame (size, amino acids)</th>
<th>Best match</th>
<th>Size (aa)</th>
<th>Organism</th>
<th>Identity (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5B2</td>
<td>2593 bp (59.58%)</td>
<td>ORF1 (93)</td>
<td>Hypothetical protein Msil_1796 (YP_002362103)</td>
<td>94</td>
<td><em>Methylocella silvestris</em> BL2</td>
<td>54/79 (68%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ORF2 (263)</td>
<td>Two-component response regulator (YP_002360627)</td>
<td>274</td>
<td><em>Methylocella silvestris</em> BL2</td>
<td>170/265 (64%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ORF3 (97)</td>
<td>Hypothetical protein Bind_3240 (YP_001831149)</td>
<td>146</td>
<td><em>Beijerinckia indica</em> subsp. indica ATCC 9039</td>
<td>23/33 (69%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ORF4* (159)</td>
<td>ECF subfamily RNA polymerase sigma-24 factor (YP_001834290)</td>
<td>190</td>
<td><em>Beijerinckia indica</em> subsp. indica ATCC 9039</td>
<td>108/143 (75%)</td>
</tr>
<tr>
<td>6S2</td>
<td>4019 bp (63.97%)</td>
<td>ORF1 (294)</td>
<td>Enoyl-CoA hydratase/isomerase (YP_549923)</td>
<td>264</td>
<td><em>Polaromonas sp.</em> JS666</td>
<td>197/255 (77%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ORF2 (617)</td>
<td>Feruloyl-CoA synthase (YP_981036)</td>
<td>618</td>
<td><em>Polaromonas naphthalenivorans</em> CJ2</td>
<td>477/618 (77%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ORF3* (306)</td>
<td>Acetyl-CoA acetyltransferase (YP_981037)</td>
<td>416</td>
<td><em>Polaromonas naphthalenivorans</em> CJ2</td>
<td>258/306 (84%)</td>
</tr>
</tbody>
</table>

* Truncated orfs are identified with an asterisk.
**Fig. 2.** Schematic representation of the genetic organization of the metagenomic inserts of p5B2 and p6S2. A bold arrow with pointed head represents a complete ORF and with a diamond head represents a truncated ORF. Transposon insertions in positive mutants are represented by (●) and in salt sensitive mutants by (▲). Thin dotted line represents pUC19 region. EcoRI/SalI fragment cloned in pRK1 and PstI fragment cloned in pRK3 are shown by dotted lines. (Note: figure not to scale.)

starvation (Mueller et al., 1992) and was later shown to be induced by many other stress conditions, including salt stress (Maul et al., 1995). Family PD027049 contain two members from E. coli K12 named YciG and YmdF. YciG and its homologue in Salmonella enterica serotype Typhimurium were shown to be regulated by RpoS (Robbe-Saule et al., 2001). RpoS is known to regulate genes important for acid, heat and salt tolerance in E. coli (Henge, 2008). Late embryogenesis abundant proteins from plants have been shown to be similar to GsiB, as they are hydrophilic proteins and contain repeats similar to GsiB (Stacy and Aalen, 1998) and they are proposed to play a function similar to GsiB under stress conditions in plants. Induction and accumulation of family PD027049 proteins under various stress conditions in different organisms point towards their role in adaptation to environmental stresses. However, exact function of these proteins is not known and they are not implicated as yet in salt tolerance. In this study, it was observed that GspM expression from a multicopy plasmid confers salt tolerance to E. coli corroborating the view that members of family PD027049 play role in stress adaptation.

3.5. Identification of gene conferring salt tolerance from p6S2

Ten salt sensitive mutants were obtained for p6S2 out of the 900 mutants screened for loss of salt tolerance ability. The transposon was located in a gene encoding for a protein of 294 amino acids in nine of these mutants and 230 bp upstream of this gene in one of the mutants. The encoded protein of about 31 kDa showed best match to a putative enoyl-CoA hydratase/isomerase from Polaromonas sp. JS666 with 77% amino acid identity. The gene was designated as echM because of its similarity with enoyl-CoA hydratase and metagenomic origin. EchM contained conserved domains for enoyl-CoA hydratase/isomerase (pfam00378), Ech), enoyl-CoA hydratase/carnitine racemase (COG1024, CaiD) and dihydroxy-naphthoic acid synthase (COG0447, MenB). Prosite search revealed presence of enoyl-CoA hydratase/isomerase signature pattern [LIVM]-[STAG]-X-[LIVM]-[DENQRHSTA]-G-X(3)-[AG](3)-X(4)-[LIVM]-[DENQRHSTA]-X-[LIVM], in between position 134–154 amino acid of the EchM as IAaLHavGGGيلةSAH1. Alignment of the EchM with rat mitochondrial enoyl-CoA hydratase (Hofstein et al., 1999; Bell et al., 2001) showed presence of conserved residues G144, E147 and E167 in EchM corresponding to rat ECH catalytic residues G141, E144 and E164 (Supplementary data).

E. coli/pRK3 showed growth similar to E. coli/p6S2 in 750 mM NaCl containing LB medium, indicating that the echM gene was responsible for the salt tolerance of the clone. E. coli/pRK4 failed to grow properly in 750 mM NaCl containing medium confirming that echM was expressed from lacZ promoter in plasmid p6S2 and pRK3 (Supplementary data).

Amino acid sequence similarity analysis with BLAST between EchM and characterized or putative enoyl-CoA hydratases from E. coli, YfgG, YfcX/FadD, PaaG, PaaF, MenB, CaiD, FadB (Yang et al., 1991) and MaoC (Park and Lee, 2003) revealed low sequence identity of the mutants. The encoded protein of about 31 kDa showed best match to a putative enoyl-CoA hydratase/isomerase from Polaromonas sp. JS666 with 77% amino acid identity. The gene was designated as echM because of its similarity with enoyl-CoA hydratase and metagenomic origin. EchM contained conserved domains for enoyl-CoA hydratase/isomerase (pfam00378), Ech), enoyl-CoA hydratase/carnitine racemase (COG1024, CaiD) and dihydroxy-naphthoic acid synthase (COG0447, MenB). Prosite search revealed presence of enoyl-CoA hydratase/isomerase signature pattern [LIVM]-[STAG]-X-[LIVM]-[DENQRHSTA]-G-X(3)-[AG](3)-X(4)-[LIVM]-[DENQRHSTA]-X-[LIVM], in between position 134–154 amino acid of the EchM as IAaLHavGGGيلةSAH1. Alignment of the EchM with rat mitochondrial enoyl-CoA hydratase (Hofstein et al., 1999; Bell et al., 2001) showed presence of conserved residues G144, E147 and E167 in EchM corresponding to rat ECH catalytic residues G141, E144 and E164 (Supplementary data).
tivity in the range of 24–35%, with maximum of 35% between EchM and FadJ. Phylogenetic analysis of EchM with its homologues formed a distinct clade separate from the clades containing E. coli enoyl-CoA hydratase proteins and their homologues (Fig. 4). EchM, an enoyl-CoA hydratase was identified to be responsible for salt tolerance ability of p6S2. Low sequence similarity and distinct evolutionary lineage of EchM in comparison to E. coli enoyl-CoA hydratases indicated that EchM may have different substrate specificity and may be performing a unique function in as yet uncharacterized pond water bacterium. Bacteria and plants have been shown to modulate their membrane lipids and fatty acids composition in response to various environmental stresses, including temperature, osmotic and salt stress (Rodriguez-Vargas et al., 2007). Salt tolerance in these cases requires removal and metabolism of these unnecessary lipids and fatty acids. Enoyl-CoA hydratases play important role in β-oxidation pathway of fatty acid metabolism. EchM might be playing a role at a rate-limiting step in lipid/fatty acid degradation process during adaptation to salt stress in E. coli. Another possible role of EchM in salt stress tolerance might be to carry out synthesis of a compatible solute such as carnitine. CaiD, a crotonobetainyl-CoA hydratase was shown to be involved in carnitine synthesis (Elssner et al., 2001). EchM has a complete conserved domain for CaiD and might be involved in shifting the metabolic flux towards carnitine synthesis.

3.6. EchM possess enoyl-CoA hydratase activity

EchM was overexpressed with a carboxy-terminal hexahistidyl tag under T7 promoter and purified by Ni²⁺-NTA affinity purification. SDS–PAGE analysis revealed that EchM was more than 95% pure after single step affinity purification (Fig 5A). Purified enzymes showed enoyl-CoA hydratase activity with crotonyl-CoA as substrate. It showed a specific activity of 5.6 U/mg with 350 µM crotonyl-CoA at pH 8.0 and 30°C.

3.7. GspM and EchM together did not result in additive response

E. coli/pRK6 was expected to express both the proteins, GspM with native promoter and EchM with lacZ promoter, whereas E. coli/pRK5 was expected to express only GspM and not EchM. E. coli/pRK6 formed smaller colonies on LB agar plates and also grew at a slower rate than the other clones and control on LB medium (Fig. 6A). Growth curve analysis of these clones in LB medium with 750 mM NaCl revealed that E. coli/pRK5 grew similar to E. coli/pRK1 but unexpectedly E. coli/pRK6 grew at a slower rate than the individual clones, E. coli/pRK1 and E. coli/pRK3 (Fig. 6B). These results indicated that expression of both the proteins together from high copy number plasmid caused stress to the E. coli culture. Optimization of appropriate expression of both the proteins will be re-

![Fig. 5. SDS–PAGE analysis of EchM. SDS–PAGE analysis of purified EchM. Lane 1, purified EchM and Lane 2, low molecular weight protein marker.](image)

![Fig. 6. Growth curves of E. coli strains. (△) E. coli/pRK1, (●) E. coli/pRK3, (▲) E. coli/pRK6, (▲) E. coli/pRK5, (●) control E. coli/pUC19 and control (○) E. coli/pLR8 in (A) LB medium and (B) in LB medium with 750 mM NaCl. Experiments were performed in triplicates, values are averages and standard deviations are shown as error bars.](image)
quired to obtain better salt tolerance from the recombinant clone expressing both the proteins.

4. Conclusion

Two novel genes involved in salt tolerance were identified from a metagenomic library prepared form pond water sample. One of the encoded proteins, GspM, shows similarity with general stress proteins and another encoded protein, EchM, shows similarity with enoyl-CoA hydratases. Enoyl-CoA hydratases are not implicated as yet in the salt tolerance process. Further studies on the novel enoyl-CoA hydratase identified in this study will help in exploring this unique mechanism of salt tolerance. These genes have potential application in generating salt tolerant recombinant bacteria or transgenic plants.

Acknowledgements

We thank Director, I.G.I.B. for constant support and encouragement. RR and AG acknowledge Council of Scientific and Industrial Research (CSIR) for Senior Research Fellowship. RKK and MP thank Department of Biotechnology, Punjabi University, Patiala, India for providing necessary facilities.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.biortech.2010.01.017.

References


Sequence analysis of a salt tolerant metagenomic clone

Raj Kishor Kapardar · Ravi Ranjan · Munish Puri · Rakesh Sharma

Received: 29 April 2010 / Accepted: 17 May 2010

Abstract Metagenome represent an unlimited resource for discovery of novel genes. Here we report, sequence analysis of a salt tolerant metagenomic clone (6B4) from a pond water metagenomic library. Clone 6B4 had an insert of 2254 bp with G+C composition of 64.06%. DNA sequence from 6B4 showed homology to DNA sequences from proteobacteria indicating origin of 6B4 metagenomic insert from a yet uncharacterized proteobacteria. Two encoded proteins from clone 6B4 showed match with ATP-dependent Clp protease adaptor protein (ClpS) and phasin, while two truncated encoded proteins showed match with poly-3-hydroxybutyrate synthase and permease. Clp complex is known to play a role in stress tolerance. Expression of ClpS from metagenomic clone is proposed to be responsible for salt tolerance of the metagenomic clone 6B4.

Keywords Metagenomic · Metagenome · Salt tolerance · ClpS

Introduction

The phenomenon of bacterial salt stress tolerance is quite complex. It has been observed that organisms which thrive in saline environmental conditions possess specific mechanisms to maintain their osmotic balance. The salt stress tolerance genes have been involved in the recovery of turgor loss due to osmotic shock by uptake and/or de novo synthesis of organic (amino acids, sugars, polyols) and inorganic (K⁺) compatible solutes and inducing general or specific stress tolerance genes [1]. Bacterial genes coding for enzymes catalyzing the synthesis of different classes of compatible solutes were successfully used for generating salt tolerant transgenic plants, for example, carbohydrate; trehalose [2], polyols; mannitol [3], amino acid derivatives; glycine betaine [4] and ectoine [5].

The microbial diversity present in the environment has been severely underestimated. This occurred profoundly since the determination of bacterial diversity relied on identifying bacteria by culturing from the environmental samples, without realizing the degree of bias introduced by the chosen culturing conditions [6]. Metagenomics has changed the way microbiologists approached different problems; it has also accelerated the rate of gene discovery. Metagenomics has vast potential which can be used in the field of biotechnology. The functional screens have identified new enzymes and antibiotics and other reagents from metagenomic libraries of diverse environments [7].

To exploit the gene pool of pond water microbial communities, a metagenomic library was prepared earlier and novel lipolytic genes were identified [8]. We envisaged that some of the clones in our metagenomic library would also possess salt tolerance genes, which may find application in
production of salt tolerant recombinant bacterial strains and transgenic plants [2, 3, 4]. To identify salt tolerant clones, we enriched the pond water metagenomic library for clones with an ability to grow in 750 mM NaCl containing medium. The enrichment resulted in isolation of three salt tolerant clones; two of the clones were sequenced and characterized earlier [1]. The third clone showed slower growth even in normal LB medium supplemented with ampicillin and hence was not characterized initially. In an attempt to identify salt tolerance genes from the third clone and to understand why the clone behaved differently from the other two clones in normal LB medium, we sequenced the third clone. Sequence analysis and gene expected to be involved in salt tolerance of this clone is presented in this report.

Clone 6B4 was found to grow at 750 mM NaCl in comparison to control which failed to grow under similar conditions [1]. The clone was further tested for its salt tolerance in LB medium supplemented with 750 mM KCl and ampicillin. The clone showed maximum growth at 28 h (Fig. 1) but the control did not show significant growth under same conditions.

The DNA sequence of both strands of p6B4 was obtained by sequencing transposon-flanking regions of the plasmids mutated by template generation system and primer walking [1]. DNA sequence of 2254 bp having G+C composition of 64.06% was obtained. The sequence obtained from the clone 6B4 was deposited in GenBank under accession number HM117648. Nucleotide BLAST analysis resulted in match in short stretches 227/264 (85% identity) with *Erythrobacter litoralis* HTCC2594, 239/287 (83% identity) with *Sphingobium japonicum* UT26S, 59/70 (84% identity) with *Bordetella avium* 197N and 30/30 (100% identity) with *Polaromonas* sp. JS666. Previously characterized salt tolerant clones from the library had a G+C composition of 59.58 and 63.97% [1]. Clone 6S2 showed significant identity with *Rhodoferax* and *Polaromonas* but clone 5B2 did not show any nucleotide identity with known sequences in the databases [1]. Even though the sequences from three clones had similar G+C composition, nucleotide BLAST analysis indicated at the possibility that metagenomic inserts in three clones originated from unrelated proteobacteria.

Four Open Reading Frames (ORFs) were identified by ORF finder. Two ORFs showed match with ClpS and phasin, while two truncated ORFs showed match with poly-3-hydroxybutyrate synthase and permease (Table 1). Organization of these ORFs is shown in Fig. 2. One of the encoded proteins showed match with 81% identity to ATP-dependent Clp protease adaptor protein, ClpS from *Erythrobacter* sp. NAP1. It had conserved domains of ClpS family (COG2127, pfam02617, PRK00033, PRK13019). Another encoded protein showed match with 55% identity to phasin from *Erythrobacter* sp. SD-21 and it had conserved domains of Phasin2 family (pfam09361, COG5490). Whereas one of the truncated encoded proteins showed match with 63% identity to poly-3-hydroxybutyrate synthase from *Erythrobacter* sp. SD-21 and it had truncated conserved domains for PhaC Poly(3-hydroxyalkanoate) synthetase (COG3243), PHA_synth_I, poly(R)-hydroxy-alkanoic acid synthase, class I (TIGR01838), and PHA_synth_II, poly(R)-hydroxyalkanoic acid synthase, class II (TIGR1839). The other truncated encoded protein showed match with 78% identity to predicted permease (YjgP/YjgQ) from *Novosphingobium aromaticivorans* DSM 12444. It had truncated domains for predicted permease YjgP/YjgQ family (pfam03739) and predicted permeases (COG0795) from permease YjgP/YjgQ family. The two truncated genes are not expected to form functional proteins and only ClpS and Phasin genes are expected to produce functional protein. These two genes may be involved in salt tolerance of the clone 6B4. Phasin proteins are important for polyhydroxy alkanoates storage and metabolism [9]. Polyhydroxy alkanoates are produced by the bacteria under environmental stress conditions [10] but are not reported to play any role in salt stress tolerance.

ClpS is a monomeric adaptor protein that binds to the N-terminal domain of ClpA subunits and alters substrate

![Fig. 1 Growth curves of *Escherichia coli* strains, (□) *E. coli* p6B4, (■) control *E. coli* pUC19 in LB medium with 750 mM KCl. Experiments were performed in triplicates, values are averages and standard deviations are shown as error bars.](image-url)
The ClpAP protease is composed of ClpA, a hexameric AAA+ ATPase/protein unfoldase, and ClpP, a 14-subunit protease [11]. ClpS directly interacts with the destabilizing N-terminal residues and transfers them to the ClpAP protease complex for degradation [12]. Clp family has been shown to act as molecular chaperones. These proteins can undo protein aggregates resulting from heat shock or other stresses and also play important role in degradation of polypeptide chains in an ATP-dependent manner [13].

ClpS is usually encoded immediately upstream of ClpA in the genomes of proteobacteria. ClpS homologs exist in bacteria. A group of ClpS is found in the genomic context of ClpA, which is universally distributed in proteobacteria and another group that is not linked to a ClpA is found broadly in actinobacteria, cyanobacteria and plant chloroplasts [14]. ClpS is distantly related to a domain of the eukaryotic E3 ubiquitin ligase, N-recognin. E3 ubiquitin ligase has been shown to be important for abiotic stress responses [15].

Salt stress and other stresses result in aggregation of cytoplasmic proteins. We propose that the multicopy expression of ClpS in metagenomic clone might be involved in increased recognition of aggregated protein for refolding or degradation by ClpAP complex and allow stress tolerance. Unwanted degradation of essential proteins due to expression of ClpS under normal growth conditions may be responsible for the slower growth of clone 6B4 in LB medium. Overexpression of the clpS gene alone under an inducible heterologous promoter will help in testing this hypothesis in near future.

**Table 1** ORF Blast results of Open Reading Frames identified from clone 6B4

<table>
<thead>
<tr>
<th>ORF</th>
<th>Protein, Organism (Acc No.)</th>
<th>Size aa</th>
<th>Identity</th>
<th>e-value</th>
<th>Family</th>
<th>CONSERVED DOMAINS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Poly-3-hydroxybutyrate synthase <em>Erythrobacter</em> sp. SD-21 (ZP_01864750)</td>
<td>626</td>
<td>61/96 (63%),</td>
<td>8e-30</td>
<td>TIGR01838</td>
<td>PHA_synth_I (Class I subfamily of poly(R)-hydroxyalkanoate synthases)</td>
</tr>
<tr>
<td></td>
<td><em>Erythrobacter</em> sp. SD-21 (ZP_01864749)</td>
<td></td>
<td></td>
<td></td>
<td>COG3243</td>
<td>PhaC Poly(3-hydroxyalkanoate) synthetase [Lipid metabolism]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>TIGR01839</td>
<td><em>PHA_synth_II</em> (Class II subfamily of poly(R)-hydroxyalkanoate synthases)</td>
</tr>
<tr>
<td>2</td>
<td>Hypothetical protein ED21_23143 <em>Erythrobacter</em> sp. SD-21 (ZP_0186479)</td>
<td>267</td>
<td>87/157 (55%)</td>
<td>4e-40</td>
<td>pfam09361</td>
<td>Phasin protein</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>COG5490</td>
<td>Uncharacterized conserved protein (Function unknown)</td>
</tr>
<tr>
<td>3</td>
<td>ATP-dependent Clp protease adaptor protein ClpS <em>Erythrobacter</em> sp. NAP1 (ZP_01040407)</td>
<td>140</td>
<td>100/122 (81%),</td>
<td>3e-52</td>
<td>pfam02617</td>
<td>Uncharacterized conserved protein (Function unknown)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>PRK00033</td>
<td>ClpS, In the bacterial cytosol, ATP-dependent protein</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>PRK13019</td>
<td>ClpS, ATP-dependent Clp protease adaptor protein ClpS</td>
</tr>
<tr>
<td>4</td>
<td>Permease YjgP/YjgQ <em>Novosphingobium aromaticivorans</em> DSM 12444 (ABD24944)</td>
<td>414</td>
<td>84/107 (78%),</td>
<td>5e-41</td>
<td>pfam03739</td>
<td>Predicted permease YjgP/YjgQ family</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>COG0795</td>
<td>Predicted permeases (General function prediction only)</td>
</tr>
</tbody>
</table>

**Fig. 2** Schematic representation of the genetic organization of the salt tolerant clone p6B4 (2254 bp), with respect to pUC19 *lacZ* promoter and *EcoRI* and *HindIII* sites of the vector. A bold arrow with a pointed head represents a complete gene and with a round head represents a truncated gene. Figure is not to scale; size of the encoded proteins is indicated in the brackets.
References


Isolation of novel lipolytic genes from uncultured bacteria of pond water

Ravi Ranjan a, Amit Grover a, Raj Kishor Kapardar b, Rakesh Sharma a,b,*

a Department of Environmental Biology, University of Delhi, Delhi 110007, India
b Institute of Genomics and Integrative Biology, Mall Road, Delhi 110007, India

Received 1 July 2005
Available online 20 July 2005

Abstract

Metagenomic libraries give access to gene pool of bacteria present in environmental samples avoiding the culture bias. A metagenomic library of pond water microbial assemblage in plasmid vector containing about 532 Mb of community DNA was prepared. Screening of a part of the unamplified library resulted in isolation of 11 unique lipolytic clones with an ability to hydrolyze tributyrin. DNA sequence of the lipolytic genes varied in G+C composition from 57% to 75%. Twelve lipolytic genes encoding proteins with 25–70% amino acid identity with proteins in the databases were identified. Ten of the encoded proteins belonged to seven known lipolytic protein families. One of the proteins was similar to recently identified esterase BioH. A lipolytic protein with high similarity to yet uncharacterized α/β hydrolase protein family abh_upf0017 was identified from one of the clones. Conserved motif for lipolytic enzymes GXSXG, conserved aspartic and histidine residues were identified in this encoded protein.

Keywords: Metagenomic; Metagenome; Lipolytic; Lipase; Esterase; α/β-hydrolase; Uncultured bacteria; Unculturable; Culture-independent; Pond water

Lipolytic enzymes include lipases (EC 3.1.1.3), which hydrolyze long chain acylglycerols (≥C10) and esterases (EC 3.1.1.1), which hydrolyze short chain acylglycerol (≤C10). Lipases are also able to hydrolyze short chain acylglycerols [1]. Lipolytic enzymes have significant biotechnological importance because of their ability to catalyze regio- and stereo-selective organic reactions [1,2]. These enzymes are also becoming important in synthetic organic chemistry because of their capability to withstand organic solvents and catalyze reverse reactions [2]. Identification of lipolytic enzymes with a spectrum of sequence and functional diversity will enrich the toolbox of synthetic chemists, which will help in quick selection of a suitable biocatalyst for challenging reactions and reaction conditions [3,4]. Large number of lipases and esterases has been identified from different domains of life. Lipase engineering database contains more than 800 protein entries assigned to 38 homologous families and 16 superfamilies [5]. Bacterial lipolytic enzymes have been classified in eight families by Arpigny and Jaeger [6]. Availability of large number of complete genome sequences is augmenting the discovery of newer lipolytic enzymes and lipolytic gene families [7–9].

Majority of the bacteria present in environment cannot be cultured in laboratory by traditional culture techniques [10]. This vast diversity of ‘yet to be cultured’ bacteria represents a large gene pool for biotechnological exploitation. Culture-independent methods like preparation of metagenomic libraries have allowed successful exploitation of microbial communities from various environmental niches [11–13]. Screening of sufficiently large metagenomic libraries can result in isolation of many genes of similar function with significant sequence variation [14,15]. Functional lipolytic genes have been previously identified from metagenomic libraries of soil [11,16,17], hot spring sediments [13], and alkaline soda lakes [18]. PCR based strategy was
used to identify a novel lipolytic gene from soil metagenome [19], which avoids the reported difficulties in functional expression of lipolytic genes in heterologous host [2].

Fresh water ponds, lakes, and water bodies have been shown to possess unique microbial diversity including members of various unculturable groups [20,21]. To the best of our knowledge, fresh water microbial communities have not been exploited by culture-independent methods for isolation of lipolytic genes. Here, we report construction of a metagenomic library from pond water microbial assemblage and identification of 11 unique lipolytic clones with an ability to hydrolyze tributyrin.

Materials and methods

Sample, strains, and culture conditions. Pond water sample was collected from University of Delhi campus in month of June 2002. It was greenish-brown in colour and had a pH of 8.5. Sample was collected in a sterile container and cells were harvested immediately for isolation of metagenomic DNA. Escherichia coli DH10B was a gift from Dr. H. Shizuya, Caltech. The E. coli strains were grown in Luria–Bertani (LB) liquid medium or on LB agar plates at 37 °C. The media were supplemented with 50 µg kanamycin and/or 100 µg ampicillin per ml for selection of plasmids.

Pond water metagenomic DNA isolation. Microbial cells from 5 L of pond water sample were harvested by centrifugation at 8000g for 10 min. Metagenomic DNA was isolated with modification of the method described by Zhou et al. [22]. The cell pellet (greenish in colour) was suspended in 20 ml of extraction buffer (100 mM Tris–HCl [pH 8.0], 100 mM EDTA, 100 mM sodium phosphate, 1.5 M NaCl, 1% cetyltrimethylammonium bromide (CTAB)) and incubated at 37 °C for 15 min. Lysozyme was added at a final concentration of 5 mg/ml and the tubes were incubated at 37 °C for 1 h with gentle shaking. Proteinase K (2.5 mg) and SDS at a final concentration of 1.0% were added, and the tubes were incubated at 65 °C for 1 h. The lysate was centrifuged at 10000g for 10 min and the supernatant was transferred to a fresh tube. It was then extracted once with equal volume of 50:50 phenol and chloroform. The DNA was precipitated from aqueous phase by 0.6 V of isopropyl alcohol and kept undisturbed for 1 h. The DNA was obtained after initial isolation procedure was resistant to digestion with restriction enzymes. The metagenomic DNA obtained after initial isolation procedure was used as described by Sambrook and Russell [25] and Ausubel et al. [23].

Transposon mutagenesis and mapping. Plasmid from the unique clones was used for in vitro transposon mutagenesis by Template Generation System (TGS, F-700) kit by the manufacturer’s provided protocol (Finzyme, Finland). The mutagenized plasmid was electroporated in E. coli DH10B and colonies containing transposon-mutated plasmids were selected on ampicillin and kanamycin containing LB agar plates. These colonies were then stabbed with the help of sterile toothpicks on 1.0% tributyrin, ampicillin, and kanamycin containing LB agar plates to select negative mutants for lipolytic activity. Fifty to 200 mutants were screened for each clone to obtain negative mutants. These mutants were expected to contain transposon insertion in lipolytic gene or in their regulatory region. Location of the transposon insertion in the sequenced mutated plasmids was mapped by size of PCR product obtained by primers against transposon ends and pUC reverse or pUC forward primer. Mutants with insertions separated by a distance of about 400–450 bp were selected for sequencing. Assembly of these sequences in most cases resulted in double strand sequence of the lipolytic gene region and single strand sequence from the ends and flanking region of lipolytic genes. Primers at the extreme ends of the single strand sequences were used to obtain double strand sequence for both the ends.

DNA sequencing and analysis. DNA sequencing reactions were done with big dye termination cycle sequencing kit version 3.1 (Applied Biosystems, USA) and the products were run on an ABI 3700 machine using POP6 method at The Centre for Genomic Application (TCGA), New Delhi, India. Transposon end specific primers SeqA and SeqB (supplied with the TGS kit) were used to sequence mutated plasmids. Sequence assembly and analysis were carried out using Lasergene package, version 5.07 (DNA Star, USA). Open reading frames in the assembled sequence for each clone were identified by ORF finder [26] at National Centre for Biotechnology Information (NCBI) website and the amino acid sequence of each identified ORF was used to find the closest match by BLAST without low complexity filter [27]. Encoded proteins were analyzed by BLAST for similarity with known lipolytic proteins and/or hydrolase fold containing proteins at Lipase Engineering Database [5] and Esther database [28], respectively. Multiple sequence alignments were carried out by ClustalW [32]. Conserved domains and patterns were analyzed using conserved protein domain database [29], InterProScan [30], and Prosite [31]. Signal peptide and transmembrane domain were predicted using server SignalP [33] and HMMTOP [34], respectively. Nucleotide sequence accession number. The sequences obtained from different clones were deposited in GenBank under Accession Nos. DQ077738–DQ077748.

Results

Construction of pond water metagenomic library

Water from a small pond in University of Delhi campus was collected. The surface of the water was covered with algal mass, which was avoided during collection of the water sample. The water appeared greenish-brown in colour and had a pH of 8.5. Microscopic observation revealed presence of bacterial, cyanobacterial, filamentous algal cells, and few possible members of invertebrate groups. Microbial cells from 5 L of pond water were harvested immediately after the collection by centrifugation. The metagenomic DNA obtained after initial isolation procedure was resistant to digestion with restriction enzymes. It was further purified by CTAB to obtain readily digestible pure DNA. CTAB is known to remove polysaccharides and humic acid impurities [22]. Pond
water metagenomic library was constructed containing more than 1,40,000 recombinant colonies. More than 90% of the colonies in the library were recombinant. Analysis of insert fragments generated by EcoRI and HindIII restriction digestion of 40 recombinant plasmids was used to estimate an average insert size of ~3.8 kb for the library (data not shown). The metagenomic library represented about 532 Mb of the pond water microbial community DNA.

**Screening of lipolytic clones**

A part of unamplified library (~30,000 colonies) was screened for lipolytic clones on 1.0% tributyrin containing LB ampicillin plates. Seventeen clones were obtained in the initial screening for their ability to produce hydrolysis zone after 48 h of incubation at 37 °C. Only 13 clones gave stable activity after single colony re-streaking. The zone size of hydrolysis of tributyrin by different clones varied from 4 to 15 mm after 72 h incubation at 37 °C, indicating variable expression or substrate preference of the lipolytic enzymes produced by the clones (Fig. 1). As no IPTG was added to the screening plates, the activity obtained was expected to be because of the expression from the native promoter in the clones or from the low level leaky expression from lac promoter. All the colonies obtained after retransformation of the plasmids isolated from the stable clones produced lipolytic activity indicating plasmid-borne nature of the lipolytic activity. Plasmids from the stable clones were designated as pLR1–pLR13. The insert size in these plasmids varied from 2.2 to 8.5 kb (Fig. 2). Restriction digestion analysis of the 13 plasmids revealed that plasmid pLR9 was a sibling of pLR11 and pLR10 of pLR13 (data not shown). Hence, pLR11 and pLR13 were not used for further analysis.

To analyze the substrate specificity of the lipolytic enzymes produced by the clones, the clones were stabbed in to 1.0% tributyrin, or olive oil and rhodamine B containing LB agar plates [35]. Most of the clones produced hydrolysis zone of varying sizes on tributyrin containing plate within 72 h of incubation at 37 °C but they failed to produce any fluorescent halo on olive oil and rhodamine B containing plates even after 7 days of incubation. These results indicated that the lipolytic enzymes produced by the clones were esterases.

**Sequence analysis**

DNA sequence of the lipolytic gene region was obtained by sequencing the transposon mutants with no lipolytic activity (Fig. 2) and primer walking. The G + C content of the sequenced region from different clones varied from 56% to 75% (data not shown) and nucleotide BLAST of these sequences did not result in any significant match. This indicated the origin of the inserts from varied bacterial phylotypes not closely related to the bacteria for which the sequence information is available in databases. DNA sequence analysis of lipolytic clones resulted in identification of candidate lipolytic gene from each plasmid. In one plasmid, pLR12, two tandem genes with similarity to lipolytic genes were identified. The proteins encoded from these genes showed 25–70% identity at amino acid level with the

---

**Fig. 1.** Lipolytic activity of different clones isolated from pond water metagenomic library on 1.0% tributyrin containing LB agar. Activity was observed after 72 h of incubation at 37 °C. Side panel represents the position of clones on the tributyrin plate, 1–13; pLR1/DH10B to pLR13/DH10B and C; pUC19/DH10B.
proteins from *Acidovorax* sp., *Azoarcus* sp., *Bacillus niacin*, *Bradyrhizobium japonicum*, *Brassica oleracea*, *Dechloromonas aromatica*, *Rubrivivax gelatinosus*, *Streptomyces coelicolor*, *Xanthomonas axonopodis*, and uncultured bacterium (Table 1). These results indicated presence of DNA from wide variety of unique bacterial phylotypes in the pond water metagenomic library.

DNA sequence of 2402 bp obtained from the selected region of the pLR1 contained an ORF, which encoded a protein of 385 amino acids. This encoded protein was most similar to a hypothetical protein blr5441 from *B. japonicum* with 52% amino acid identity. It had a complete domain for \(\beta\)-lactamases. Close matches of lipolytic protein from pLR1 contained \(\beta\)-lactamase domain and many of them were annotated as putative \(\beta\)-lactamases or penicillin-binding proteins in the database. We found no significant match for this protein in lipase engineering database but comparison with bacterial lipolytic enzymes revealed that it belonged to \(\beta\)-lactamase fold containing bacterial esterase family VIII of lipolytic enzymes [6]. Multiple sequence alignment of protein from pLR1 with members of family VIII revealed the presence of conserved \(\beta\)-lactamase SXXK motif and another conserved motif LLXHXXG in the encoded protein (Fig. 3A).

An encoded protein of 277 amino acids from pLR2 showed maximum similarity with chlorophyllase 2 from *B. oleracea* with 25% amino acid identity. The other

Table 1

<table>
<thead>
<tr>
<th>Clone</th>
<th>ORF GC (%)</th>
<th>ORF (aa)</th>
<th>Best match (Accession No.)</th>
<th>Organism</th>
<th>e value</th>
<th>Identity (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LR1</td>
<td>67.36</td>
<td>385</td>
<td>Hypothetical protein (NP_772081)</td>
<td><em>B. japonicum</em></td>
<td>1e-107</td>
<td>204/386 (52)</td>
</tr>
<tr>
<td>LR2</td>
<td>72.66</td>
<td>277</td>
<td>Chlorophyllase 2 (AAN51934)</td>
<td><em>B. oleracea</em></td>
<td>2e-04</td>
<td>38/152 (25)</td>
</tr>
<tr>
<td>LR3</td>
<td>71.52</td>
<td>315</td>
<td>Lipase/esterase (AAS77247)</td>
<td>Uncultured bacterium</td>
<td>3e-71</td>
<td>148/312 (47)</td>
</tr>
<tr>
<td>LR4</td>
<td>71.07</td>
<td>294</td>
<td>(\delta(\beta)-)Hydroxybutyrate oligomer hydrolase (BAB19271)</td>
<td><em>Acidovorax</em> sp.</td>
<td>1e-116</td>
<td>203/290 (70)</td>
</tr>
<tr>
<td>LR5</td>
<td>57.71</td>
<td>278</td>
<td>Predicted hydrolases or acyltransferases (YP_158932)</td>
<td><em>Azorarcus</em> sp.</td>
<td>3e-68</td>
<td>131/255 (51)</td>
</tr>
<tr>
<td>LR6</td>
<td>75.08</td>
<td>304</td>
<td>Esterase/lipase (ZP_00245626)</td>
<td><em>R. gelatinosus</em></td>
<td>4e-92</td>
<td>200/286 (69)</td>
</tr>
<tr>
<td>LR7</td>
<td>64.05</td>
<td>343</td>
<td>Putative hydrolase (CAC01373)</td>
<td><em>S. coelicolor</em></td>
<td>7e-45</td>
<td>121/287 (42)</td>
</tr>
<tr>
<td>LR8</td>
<td>61.73</td>
<td>458</td>
<td>Esterase S4 (AAQ03995)</td>
<td><em>B. niacin</em></td>
<td>2e-77</td>
<td>183/473 (38)</td>
</tr>
<tr>
<td>LR9</td>
<td>76.15</td>
<td>259</td>
<td>Biotin biosynthesis protein (AAM35277)</td>
<td><em>X. axonopodis</em></td>
<td>3e-41</td>
<td>114/242 (47)</td>
</tr>
<tr>
<td>LR10</td>
<td>61.49</td>
<td>334</td>
<td>Predicted hydrolase (ZP_00348574)</td>
<td><em>D. aromatica</em></td>
<td>7e-92</td>
<td>177/317 (55)</td>
</tr>
<tr>
<td>LR12A</td>
<td>57.73</td>
<td>301</td>
<td>Predicted hydrolases or acyltransferases (ZP_00243877)</td>
<td><em>R. gelatinosus</em></td>
<td>2e-78</td>
<td>148/295 (50)</td>
</tr>
<tr>
<td>LR12B</td>
<td>57.06</td>
<td>287</td>
<td>Predicted esterase (ZP_00152189)</td>
<td><em>D. aromatica</em></td>
<td>1e-75</td>
<td>144/255 (56)</td>
</tr>
</tbody>
</table>

\(a\) aa, amino acids.
close matches of this protein were chlorophyllase, predicted dienelactone hydrolase, and lipase from Arabidopsis thaliana, Nostoc punctiforme, and S. coelicolor, respectively. This protein belonged to Moraxella lipase 1 family of lipolytic enzymes (Fig. 3B). Signal peptide and a single transmembrane region from 5th to 22nd amino acid at amino-terminal region of encoded lipolytic protein from pLR2 were predicted indicating that it might be a membrane bound enzyme.

DNA sequence of 2453 bp obtained from the selected region of the pLR3 resulted in identification of an ORF encoding for a protein of 444 amino acids. Homology search revealed best match of this protein with a lipase/esterase from uncultured bacterium of 311 amino acids. Most of the other close matches were also similar in length. Analysis of the sequence for start sites downstream of the predicted start site helped in identification of two more putative start sites at 901 and 1144 bp, which form ORFs encoding for proteins of 396 and 315 amino acids, respectively. A putative ribosome-binding site (RBS) AGGAGA was detected upstream of start codon at 1144 bp. Presence of RBS and matches of similar size proteins in other genomes strongly indicated that the start codon at 1144 bp might be the actual start codon.
Fig. 4. Conserved sequence blocks from multiple sequence alignment of lipolytic proteins from pond water metagenomic library. (A) LR5 # DQ077742 with members of non-heme peroxidases lipolytic proteins: Atu, Agrobacterium tumefaciens type strain C58 # AAK89901; Bpy, Burkholderia pyrocinia # P25026; Rle, Rhizobium leguminosarum bv. trifolii # AAL17798; Ppu, Pseudomonas putida # BAA86922; Pae, P. aeruginosa PAO1 # G83024; Rec, Rhodococcus erythropolis # DQ077747 with members of family abh_upf0017 (Esterase database): Rso, R. solanacearum # AL049862; Hsa, Homo sapiens # AAL17798; Ppu, # AY033290. (B) LR8 # DQ077745 with members of family Bacillus esterases lipolytic proteins: Sav, Bacillus subtilis sp. with 70% amino acid identity. This protein belonged to family Moraxella lipase 2 family of hormone sensitive lipase superfamily (Fig. 3C). InterProScan search revealed the presence of HTH_FIS domain at the carboxyl-terminal of this protein. We

start site for the lipolytic protein expressed from pLR3. This protein belonged to family Moraxella lipase 2 of hormone sensitive lipases superfamily (Fig. 3C).

Analysis of 1201 bp DNA sequence of pLR4 revealed the presence of an ORF encoding for a protein of 294 amino acid with high similarity to \( d(-) \)-3-hydroxy-butyrate oligomer hydrolyase of \textit{Acidovorax} sp. with 70% amino acid identity. This protein belonged to family Moraxella lipase 3 of superfamily non-heme peroxidases (Fig. 3D).

A predicted ORF encoding for 198 amino acid long protein from pLR5 showed best match with predicted hydrolase or acyltransferase from \textit{Azorarcus} sp. of 289 amino acids with 51% identity. Search for an alternate start site revealed the presence of a TTG start site 240 bp upstream of the observed start site, coding for a protein of 278 amino acids. A putative ribosome-binding site AAGGAA was observed upstream of this TTG start site indicating that this might be the actual start site of the encoded protein from pLR5. Search in prosite database with encoded lipolytic protein from pLR5 revealed the presence of ribosomal protein S7 signatures [DENSK]-X-[LIVMDET]-X(3)-[LIVMFTA] \( \times(6) \)-G-K-[KR]-X(5)-[LIVMF]-[LIVMFC]-X(2)-[STAC] at amino-terminal region. This pattern was absent in the close protein matches of lipolytic protein from pLR5 obtained by BLAST in NCBI database. This protein belonged to family non-heme peroxidase of superfamily non-heme peroxidase (Fig. 4A).

A gene encoding 304 amino acid long protein showed best match with putative esterase/lipase of \textit{R. gelatinosus} with 68% identity at amino acid level and was identified from 1437 bp DNA sequence obtained from pLR6. This protein belonged to Moraxella lipase 2 family of hormone sensitive lipase superfamily (Fig. 3C). InterProScan search revealed the presence of HTH_FIS domain at the carboxyl-terminal of this protein. We
failed to find similar domain in other close matches of this protein from the databases.

Lipolytic protein of 343 amino acids was identified from 1622 bp DNA sequence obtained for the pLR7. It had maximum homology with putative hydrolase from S. coelicolor with 42% amino acid identity. This protein contained an α/β hydrolase conserved domain and belonged to family Moraxella lipase 3 of non-heme peroxidase superfamily of lipolytic proteins (Fig. 3D).

Identified protein from pLR8 showed maximum homology with Esterase 54 of B. niacin with 38% amino acid identity. This encoded protein of 458 amino acids belonged to Bacillus esterase family of carboxylesterase superfamily of lipolytic enzymes (Fig. 4B). Analysis in prosite database revealed carboxylesterase type B serine active site signature1 {F-[GR]-G-X(4)-[LIVM]-X-[LIV]-X-G-X-S-[STAG]-G} in amino acid sequence of encoded protein from pLR8.

An ORF of 227 amino acids from pLR9 in the lipolytic gene region matched with biotin biosynthesis protein, BioH of X. axonopodis with 47% amino acid identity. Comparison of encoded protein with other BioH proteins revealed truncation at N-terminal region of the encoded protein. Search for alternate start sites revealed the presence of a TTG and a GTG initiation codon at 60 and 96 bp upstream of the identified ATG codon. A putative RBS was identified upstream of the GTG codon indicating that the GTG codon might be the actual initiation codon for the lipolytic protein encoded from pLR9. The bioH gene from E. coli has been shown to possess esterase activity after the structure analysis revealed the presence of catalytic triad and active site similar to lipolytic enzymes [36].

The DNA sequence from the lipolytic gene region of pLR10 encoded a protein of 334 amino acids with best match to a predicted hydrolase of D. aromatica with 55% amino acid identity. Conserved domain search at NCBI and InterProScan revealed the presence of predicted α/β hydrolase domain in the encoded protein. No significant match was found in lipase engineering database for this protein. Search in Esther database revealed maximum match of this protein with members of uncharacterized α/β hydrolase protein family abh_upf0017. Protein members from bacteria, yeast, plants, and animals represent this uncharacterized family. Prosite database contains a signature pattern [D-X(8)-[GH]-[LYF]-X(4)-[DET]-[LY]-Y-X(3)-[ST]-X(7)-[IV]-X(2)-[PS]-X-[LIVM]-X-[LIVM]-X(3)-[DN]-D] for this family. Lipolytic protein from pLR10 contained this conserved pattern with only one amino acid difference. These proteins along with encoded protein from pLR10 contain the characteristic catalytic triad active site serine motif GXXSG, conserved aspartic acid, and histidine residues of lipolytic enzymes (Fig. 4C). These results indicate that encoded protein from pLR10 is a new member of yet uncharacterized α/β hydrolase protein family and members of this family are expected to possess lipolytic activity.

DNA sequence obtained from pLR12 contained two tandem ORFs similar to lipolytic genes. We obtained transposon mutants with no lipolytic activity located in both the identified probable lipolytic genes in pLR12; hence, it is difficult to predict if the activity loss is because of disruption of the individual gene or because of the polarity effect. As both the genes show significant similarity with lipolytic proteins, we assume that both the identified genes may possess lipolytic activity. One of the ORF encoded for a 301 amino acid long protein, which showed maximum homology to predicted hydrolase or acyltransferase of R. gelatinosus with 50% amino acid identity. The encoded protein belonged to family Moraxella lipase 3 of non-heme peroxidase superfamily (Fig. 3D). The second ORF encoded a protein of 287 amino acids, which was most similar to predicted esterase of D. aromatica with 56% amino acid identity. Conserved domain search revealed the presence of complete patatin and esterase domains. The patatin domain containing bacterial lipolytic proteins form a newly described probable bacterial lipolytic protein family [8]. This encoded protein was predicted to have a signal peptide and a transmembrane region from 7th to 29th amino acid at amino terminal.

Discussion

Majority of bacteria in environment remain unculturable by the routine laboratory cultivation methods. Culture-independent microbial diversity analysis revealed the presence of unculturable bacteria in most environmental samples. Half of the known phyla in eubacteria are not represented by any cultured member [37], indicating the vast phylogenetic diversity of these “yet to be cultured” bacteria. Metagenomic libraries have been exploited for identification of novel genes for bioactive compounds and biocatalysts [11,12,14,38] avoiding the culture bias. Fresh water samples have been shown to possess high microbial diversity including members of unculturable clades. We have constructed a metagenomic library of pond water microbial assemblage containing about 532 Mb of community DNA. A part of this library was screened for isolation of 11 unique lipolytic clones, which is significantly higher in comparison to the lipolytic clones isolated from soil libraries [16,17]. Isolation of low number of functional lipolytic clones can be attributed to probable non-recognition of regulatory elements from unculturable bacteria in E. coli and reported difficulties in expression of lipolytic genes in heterologous hosts because of requirement of folding or transport partners [2] or because of toxicity of expressed lipolytic protein [19]. Isolation of higher number of functional lipolytic clones from our sample may be
because of our careful and deliberate attempt to select clones even with very low lipolytic activity, reasoning that many of the novel lipolytic genes may not express well in *E. coli* because of the above-mentioned factors. The other reason can be the difference in the composition of the bacterial assemblage in pond water sample in comparison to soil samples used in the other studies.

Significant G+C composition difference varying from 56% to 75% was observed in the DNA sequence of lipolytic gene containing region of the clones. In addition, the encoded proteins matched with 25–70% amino acid identity with proteins from bacteria belonging to proteobacterial, actinobacterial, and firmicute groups. This indicated that the DNA belonging to various phylogenotypes representing diversity of pond water microbial assemblage was cloned in the metagenomic library. This is in good correlation with other studies demonstrating expression of genes from various metagenomic samples in *E. coli* [11,12,14,15], indicating suitability of *E. coli* as a host for metagenomic studies. However, the use of other hosts is expected to increase the discovery rate of the functional clones.

Lipolytic clones identified from the pond water metagenomic library differed significantly in their ability to hydrolyze tributyrin in agar plates. There can be many possible reasons for this variable activity of the clones; one could be the difference in expression level of the different cloned genes. The G+C contents of the sequenced clones vary significantly and in many cases, we failed to identify the putative ribosome-binding site near to putative start site, which reflect that yet uncharacterized bacteria in pond water microbial assemblage differ in codon usage and regulatory elements from *E. coli*.

In one of the clones, pLR5, a TTG and in another, pLR9, GTG codons were identified as putative initiation site. Alternate initiation codons TTG and GTG are known to generally result in lower expression than ATG initiation codon in *E. coli* [39]. Other possible reason could be the intracellular or membrane bound (LR2 and LR12B) nature of the encoded lipolytic proteins, which will not allow enzyme to diffuse in the plate and would only allow formation of hydrolysis zone near to the colony. Another reason could be the difference in the substrate preference of the encoded enzymes as each of the identified protein differed significantly from each other in amino acid sequence. Ten of the identified proteins belonged to seven known lipolytic protein families, one of the proteins matched with BioH and other with yet uncharacterized α/β hydrolase family. Members of different families of lipolytic enzymes have been shown to vary in their substrate preference. The substrate used in the plate assay may not be the most preferred substrate for the lipolytic enzymes produced by the clones isolated from pond water metagenomic library.

In this study and in a previous functional metagenomic study [18] lipolytic proteins similar to BioH were identified. Recently function for this protein of biotin biosynthesis operon was identified by structural genomics approach as an esterase [36]. Structural analysis of BioH protein from *E. coli* revealed the presence of a catalytic triad similar to lipolytic enzymes, which prompted study on the catalytic activity of this protein. Encoded protein from pLR10 showed high homology to proteins belonging to family abh_upf0017 in ESTher database. Our results indicate that yet uncharacterized abh_upf0017 family members are expected to possess lipolytic activity. Identification of BioH like proteins and a novel lipolytic protein with high similarity to members of uncharacterized abh_upf0017 protein family reflects the potential of functional metagenomic studies in identification of function for many hypothetical and unknown function proteins even from the genome sequences of well-characterized culturable bacteria.

In conclusion, we have prepared a metagenomic library from pond water microbial assemblage containing more than 532 Mb of community DNA. Twelve unique lipolytic genes were identified with low similarity to known lipolytic proteins. Ten of the identified lipolytic proteins represented seven known lipolytic enzyme families. An encoded protein was most similar to recently identified esterase BioH. One novel lipolytic protein with low homology to known lipolytic proteins, belonging to uncharacterized α/β hydrolase family, was identified. They are expected to have different substrate specificity, which will be useful in developing them as biocatalysts for synthetic organic chemistry. Overexpression and purification of the proteins identified in this study are in progress to further characterize their substrate preference and biotechnological potential.

Acknowledgments

We thank Prof. Samir K. Brahmachari, Director, I.G.I.B. for encouragement and support. R.S. thanks Prof. Rup Lal for allowing use of his laboratory facilities in initial phase of this work. R.R. and A.G. acknowledge CSIR for pre-doctoral research fellowship. This work was supported by Grant SR/FTP/LS-277/2000 from Department of Science and Technology.

References


‘Unculturables’ bacterial diversity: An untapped resource

Rakesh Sharma1,2,* , Ravi Ranjan2, Raj Kishor Kapardar1 and Amit Grover2

1Environmental Biotechnology Unit, Institute of Genomics and Integrative Biology, Mall Road, New Delhi 110 007, India
2Department of Environmental Biology, University of Delhi, Delhi 110 007, India

More than 99% of bacteria from environmental samples remain ‘unculturables’ in the laboratory. Many of these ‘unculturables’ bacteria represent new phylotypes, families and divisions in domain bacteria and archaea. ‘Unculturables’ bacterial diversity presents a vast gene pool for biotechnological exploitation and poses a major challenge for microbiologists to understand their phylogenetic relationship and ecological significance. New culture techniques along with culture-independent methods, like PCR amplification from microbial community DNA (metagenome) and functional or sequence-based screening of metagenomic DNA libraries are proving useful for exploitation of ‘unculturables’ bacteria. Environmental genome sequencing efforts are aiding better understanding of the community structure and their physiological importance in the ecosystem.

HUMAN pursuit has led to cultivation of bacteria from almost all possible habitats on earth for studies related to bacterial diversity, diseases, ecological functions and biotechnological applications. Bacteria were initially isolated from habitats commonly associated with humans, at near neutral pH and ambient temperature. Later these were isolated from even the most hostile environments, like thermal vents, acidic ponds, saturated brine and glaciers. Successful isolation of bacteria from earth’s crust and polar ice has led to the belief that bacterial life may exist even on other planets. Though the search to find bacterial life on other planets is ongoing, we are still far away from understanding the bacterial diversity in even the most common and well-studied niches.

Bacterial diversity, the ‘unculturables’

Studies in the last two decades have revealed that most (more than 99%) of the bacteria present in many environmental samples cannot be cultivated in the laboratory and hence remain obscure for their ecological functions, and unexploited for biotechnological applications. Culture-independent studies (mostly based on 16s rRNA gene sequence analysis) have made it apparent that a large proportion of these ‘yet to be cultivated’ bacteria belong to new genotypes, classes and divisions in the domains eubacteria and archaea. These studies have revealed presence of newer bacteria even in samples thought to be most well characterized like dental plaques, sea water and garden soil. Molecular analysis of bacterial diversity has already resulted in identification of 40 divisions in eubacteria, including 13 candidate divisions, which are not represented by any cultured member. These studies have also revealed presence of archaea in habitats like soil, seawater, etc., contrary to previous belief that archaea inhabited only the extreme environments.

Cultivating the ‘unculturables’

Abundance and diversity of unculturable bacteria in almost all environmental niches have led to the understanding that the so-called ‘unculturables’ bacteria actually multiply in their natural environment and if suitable culture conditions were provided it should be possible to cultivate them in the laboratory. In the earliest cultivation attempts, media with very low nutrient were used considering the high nutrient contents of common laboratory media as compared to those present in the natural environments. Successful isolation of many new genera by these methods including members of the candidatus clade like SAR11 resulted in more efforts in this direction. Recently, using simple media and physiological conditions, many previously unknown bacteria have been isolated which belong to diverse families and phyla. Extinction culturing technique with low nutrient media was used to culture previously uncultured marine bacteria belonging to SAR11, OM43, SAR92 and OM60/OM241 clades of proteobacteria. A collection of 350 isolates from soil was obtained using a variety of simple solid media in petri dishes. Twenty seven per cent of these isolates belonged to 20 as yet unnamed family level groupings. Many of these isolates were members of poorly studied subdivisions of phyla Acidobacteria, Verrucomicrobia, Gemmatimonadetes and Actinobacteria. Recently, bacteria from soil and termite gut were isolated by an integrated approach using various growth parameters like low nutrient media, varying oxygen and carbon dioxide concentrations, long incubation period and additives like humic acids and quorum signalling molecules. It was observed that more isolates belonging to Acidobacteria were obtained in presence of higher concentration of carbon dioxide.

*For correspondence. (e-mail: rsharma@igib.res.in)
Other innovative and successful approaches considered include, simulation of natural environment, community interactions and cell–cell communication important for the cultivation. Kaeberlein et al.\textsuperscript{17} designed a diffusion chamber that allowed cultivation of previously uncultivated bacteria in a simulated natural environment. The isolates did not grow on artificial media alone but formed colonies in presence of other microorganisms indicating that they required specific signals originating from their neighbours that point to the presence of familiar environment. In another method, single cells were encapsulated in gel microdroplets and allowed to grow with nutrients that were present at environmental concentrations in a single vessel. The pore size in gel microdroplets allowed exchange of metabolites and signalling molecules, which might be the reason for the enhanced culturability\textsuperscript{18}.

**Mining the ‘unculturables’**

Culture-independent molecular microbial diversity studies not only revealed the vast unexplored bacterial diversity but also helped in development of techniques suitable for isolation of high purity microbial community DNA (metagenome) from various environmental samples. The metagenomic DNA gave molecular biologists a chance to peek into the genomes of ‘unculturable’ bacteria, circumventing the need to culture them in the laboratory. The possibility of getting access to the huge and diverse gene pool of ‘unculturable’ bacteria led to a ‘gold rush’ in the last decade, with publications describing innovative techniques to exploit metagenomes for novel biocatalysts and bioactive compounds (Table 1). The metagenomes from many habitats have been exploited for isolation of novel genes by PCR amplification using primers against the conserved domains of known genes or by preparation of metagenomic libraries containing small inserts of 2–15 kb in plasmid vectors or large inserts of 40–130 kb in cosmid, fosmid or bacterial artificial chromosome (BAC) vectors (Figure 1).

**PCR cloning**

Two β-ketoacyl genes, components of bacterial type II polyketide synthases, were cloned from environmental DNA by PCR amplification using primers against the conserved domain of known ketosynthase and acyl carrier protein (ACP) genes\textsuperscript{19}. PCR amplification using primers against conserved domains and genome walking resulted in isolation of two complete genes for 2,5-diketo-d-gluconic acid reductase. Compared to previously known 2,5-diketo-D-gluconic acid reductases, they had some valuable properties like lower $K_m$ values and higher thermostability\textsuperscript{20}. PCR method was also used to explore the diversity of chitinase genes in culturable and unculturable marine bacteria\textsuperscript{21}. A lipase gene with less than 20% similarity to known lipases at amino acid level was cloned from soil DNA using primers against the conserved domains of the known lipases and genome walking to obtain the full length gene\textsuperscript{22}. A major disadvantage of PCR-based method is that the sequence information of known genes from culturable bacteria is used, thus eluding the possibility of finding completely novel genes. Gene cassette PCR method was devised using 59 base element family of recombination sites as target. These sites flank gene cassettes associated with integrons. This method allowed sequence-independent recovery of entire genes from environmental DNA. Genes with different possible functions were amplified using this method from soil metagenome\textsuperscript{23}.

<table>
<thead>
<tr>
<th>Biocatalyst and bioactive compound</th>
<th>Vector</th>
<th>Environmental sample</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agarase</td>
<td>Cosmid</td>
<td>Soil</td>
<td>43</td>
</tr>
<tr>
<td>Alcohol oxidoreductase</td>
<td>Plasmid</td>
<td>Soil, sediment</td>
<td>28</td>
</tr>
<tr>
<td>Amidase</td>
<td>Cosmid</td>
<td>Soil</td>
<td>43</td>
</tr>
<tr>
<td>Amylase</td>
<td>Cosmid, BAC</td>
<td>Various samples, soil</td>
<td>27, 36, 43</td>
</tr>
<tr>
<td>Cellulase</td>
<td>Cosmid</td>
<td>Soil</td>
<td>43</td>
</tr>
<tr>
<td>Chitinase</td>
<td>Lambda-ZAP</td>
<td>Marine water</td>
<td>24</td>
</tr>
<tr>
<td>DNase</td>
<td>BAC</td>
<td>Soil</td>
<td>36</td>
</tr>
<tr>
<td>Esterase</td>
<td>Plasmid</td>
<td>Soil</td>
<td>26</td>
</tr>
<tr>
<td>Glycerol/diol dehydratase</td>
<td>Plasmid</td>
<td>Soil, sediment</td>
<td>32</td>
</tr>
<tr>
<td>4-Hydroxybutyrate dehydrogenase</td>
<td>Plasmid</td>
<td>Soil</td>
<td>25</td>
</tr>
<tr>
<td>Lipase</td>
<td>Plasmid, cosmid, BAC</td>
<td>Soil</td>
<td>22, 26, 36, 43</td>
</tr>
<tr>
<td>Nitrilase</td>
<td>Lambda-ZAP</td>
<td>Various samples</td>
<td>29</td>
</tr>
<tr>
<td>Pectate lyase</td>
<td>Cosmid</td>
<td>Soil</td>
<td>43</td>
</tr>
<tr>
<td>Xylanase</td>
<td>Lambda-ZAP</td>
<td>Insects gut</td>
<td>30</td>
</tr>
<tr>
<td>Indirubin</td>
<td>BAC</td>
<td>Soil</td>
<td>40</td>
</tr>
<tr>
<td>Long chain N-acyl aromatic amino acids</td>
<td>Cosmid</td>
<td>Soil</td>
<td>38</td>
</tr>
<tr>
<td>Polyketide synthase</td>
<td>Cosmid</td>
<td>Soil</td>
<td>42</td>
</tr>
<tr>
<td>Turbomycin</td>
<td>BAC</td>
<td>Soil</td>
<td>41</td>
</tr>
<tr>
<td>Violacein, deoxyviolacein</td>
<td>Cosmid</td>
<td>Soil</td>
<td>38</td>
</tr>
</tbody>
</table>
Plasmid cloning

Small insert metagenomic libraries were prepared using Escherichia coli as host for various environmental samples including marine water, sediments and soil from various regions. These libraries were used for functional screening of clones expressing various enzymes like chitinase, hydroxybutyrate dehydrogenase, lipase and esterase, amyrase, alcohol oxidoreductase, nitrilase and xylanase. Two libraries prepared from coastal and estuarine waters were screened for clones expressing chitinase. Nine clones were obtained from 75,000 colonies screened from estuarine sample and one clone for every 500 recombinants was tested from coastal library. The results were consistent with the culture-based estimates of the proportion of marine bacteria that degrade chitin. Three different soil libraries were constructed and screened for various enzymatic activities. These libraries containing more than 930,000 recombinants were screened to obtain five clones that grow with 4-hydroxybutyrate as sole carbon and energy source. Gene sequences from three clones showed similarity to known genes of 4-hydroxybutyrate dehydrogenase, member of DedA family and enoyl coenzymeA hydratase/isomerase. The other two sequenced inserts showed no homology to known genes in databases. One lipase and three esterase clones were obtained by screening the three soil libraries. The deduced protein sequences of these four clones showed less than 50% amino acid identity to known sequences in databases, indicating that sequence information collected solely with cultivated microorganisms was not sufficient to design universal primers to retrieve the variety of genes encoding lipolytic enzymes from natural microbial communities. Utilizing the same three soil libraries, Majernik et al. complemented the Na⁺/H⁺ antiporter deficient E. coli strain. They obtained two clones, one containing a novel Na⁺/H⁺ antiporter belonging to NhaA family and the other with the DNA region of E. coli K12 containing nhaA, nhaR and gef. A novel amyrase with an exceptional process compatibility and economics was obtained by a combination of environmental library screening and laboratory evolution approaches. Screening of >600-biototype specific environmental DNA libraries containing 10⁶ to 10⁹ members per library by selection-based expression assay resulted in isolation of 137 nitrilases. Sequence analysis revealed five major sequence clades within the nitrilase subfamily. Libraries prepared from the intestinal tracts of termites and moths were screened to obtain unusual microbial xylanases. Biochemical analysis revealed that they were true xylanases though their sequences were remarkably different from the known xylanases.

In another technique which combined traditional enrichment culture and metagenomic approach, a gene bank of enteric bacteria was generated containing 24 clones. These clones were completely sequenced. Sixteen of them conferred carbonyl forming phenotype and eight clones exhibited NAD(H)-dependent alcohol oxidoreductase activity. Libraries from three different environmental samples were prepared after enrichment culture with glycerol under anaerobic conditions and screened by functional assay and colony hybridization to obtain glycerol dehydratase and diol dehydratase encoding genes. Out of the seven clones obtained, two clones were similar to glycerol dehydratase of Citrobacter freundii and the other five contained dehydratase encoding gene region similar to enteric bacteria.

Cosmid, fosmid or bacterial artificial chromosome cloning

Large insert libraries are particularly suitable for the screening of operons and pathways for the novel bioactive compounds. Cosmid or fosmid libraries containing inserts of about 40 kb and bacterial artificial chromosome (BAC) libraries containing inserts up to 150 kb have been prepared from metagenomic DNA. The major difficulty in preparation of large insert libraries is isolation of high molecular weight metagenomic DNA, which is a challenging task for samples rich in organic components, humic
acids and heavy metals. This could be the reason for early success in preparation of large insert metagenomic libraries from marine samples. A fosmid library containing inserts of ~40 kb marine bacterial DNA was constructed by Stein et al., followed by preparation of BAC library of marine microbial assemblages containing inserts up to 150 kb with an average insert size of 80 kb. Rondon et al. were the first to prepare large insert BAC libraries from soil metagenomic DNA for screening of antibiotics and biocatalytic activities. Two metagenomic BAC libraries, having average inserts size of 27 kb and 44.5 kb, containing more than 1 Gbp of soil DNA were prepared. Phylogenetic analysis of 16S rRNA genes amplified from these libraries indicated presence of DNA of wide diversity of microbial phyla. Active clones with antibacterial, lipase, amylase, nuclease and hemolytic activity were identified from these two libraries. Isolation of long chain acyltyrosine antibiotics, derived from a single environmental DNA open reading frame, were reported in another study, followed by identification of a biosynthetic gene cluster that produced two additional families of natural products containing more than 1 Gbp of soil DNA were prepared. Phylogenetic analysis of 16S rRNA genes amplified from these libraries indicated presence of DNA of wide diversity of microbial phyla. Active clones with antibacterial, lipase, amylase, nuclease and hemolytic activity were identified from these two libraries. Isolation of long chain acyltyrosine antibiotics, derived from a single environmental DNA open reading frame, were reported in another study, followed by identification of a biosynthetic gene cluster that produced two additional families of natural products derived from long chain N-acetyltyrosines.

A blue clone was isolated and characterized from environmental DNA cosmid library. Colour-producing clones often indicate presence of small molecules. Characterization of blue clone revealed presence of a four-gene biosynthetic cluster conferring production of violacin and deoxyviolacin in E. coli host. The DNA sequence of the gene cluster resembled with that of violacin gene cluster sequenced from cultured bacterium Chromobacterium violaceum. A BAC soil DNA library, containing inserts between 5 and 120 kb resulted in isolation of many antibacterial clones. One of these clones contained indirubin and related small molecules. In another study, characterization of dark brown colour producing BAC metagenomic clones revealed presence of broad spectrum antibiotic compounds Turbomycin A and Turbomycin B. Sequence analysis of the clones revealed that a single open reading frame was necessary and sufficient to confer brown, orange and red pigments on E. coli. It was demonstrated that interaction of indole with homogentisic acid, synthesized by 4-hydroxyphenylpyruvate dioxygenase encoded by open reading frame from the BAC, resulted in production of Turbomycin A and Turbomycin B in recombinant clone.

A large number of antibiotics and bioactive compounds are synthesized by actinomycetes in nature. It was realized that it would be advantageous to prepare metagenomic libraries in Streptomyces as it will help in better expression of the genes from related unculturable actinomycetes. A cosmid library prepared in shuttle cosmid vector was transformed in Streptomyces lividans and screened for polyketide synthase genes by PCR and presence of novel molecules by high performance liquid chromatography. Eight new polyketide synthase genes and five clones producing new molecules were identified. Voget et al. utilized the precultivation-metagenomic technology for preparation of cosmid library from a microbial consortium. This library was successfully used for prospecting of various biocatalysts like agarase, amylase, cellulase, pectate lyase and lipase. In an attempt to utilize the expression capabilities of different hosts like E. coli, Streptomyces lividans and Pseudomonas putida new tools were developed. These tools included a new BAC vector capable of transferring large fragments of environmental DNA among these three hosts, modified S. lividans and P. putida strains and high throughput BAC library transfer methods. The development of new tools exploiting expression capabilities of multiple hosts, libraries prepared from metagenomic DNA from various habitats, in combination with high throughput screening is expected to yield novel and potent bioactive compounds in near future.

Understanding the ‘unculturables’

One to two orders of magnitude higher populations count by microscopic observation than that measured by conventional culturing of the same sample indicated presence of a large number of ‘unculturable’ bacteria in the environment. Microbial diversity analysis by 16S rRNA gene sequences confirmed their presence and diverse phylogenetic linkages but this method failed to elucidate their physiology, abundance and ecological significance. Attempts in this direction would help in understanding ecosystem function, community structure and also in designing newer techniques to cultivate the ‘unculturable’ bacteria. Genomic information of the ‘unculturables’ bacteria can be obtained from the clones in metagenomic libraries, which can help in understanding their physiology and also about their role in ecosystem. Some success has been obtained in achieving these goals in recent years. Sequence analysis of a 130 kb genomic fragment from a BAC clone that encoded the rRNA operon from an uncultivated member of marine γ-proteobacteria (SAR 86 group) revealed the presence of rhodopsin gene. This gene encoded a protein with highest amino acid sequence similarity with archaean rhodopsins. The halophilic archaean contains bacteriorhodopsins that function as light-driven proton pump for energy generation. This first report of the presence of rhodopsin in marine γ-proteobacteria in domain bacteria suggested the possibility of a previously unrecognized phototrophic pathway that may influence the flux of carbon and energy in the ocean’s photozone worldwide.

The contents and structural comparison of 34 kb fragment containing 16S/23S rRNA operon and 17 genes encoding putative proteins from a non-thermophilic clade of soil crenarchaeota (archaea) revealed significant differences from their previously studied marine relative. In order to understand the physiology of the yet to be cultured members of the Acidobacteria division, a BAC clone containing full rRNA gene and 20 other open reading
frames was completely sequenced. Genes for cell division, cell cycling, folic acid biosynthesis, substrate metabolism, amino acid uptake, DNA repair and transcriptional regulation were identified. Advancement in sequencing technology, reduction in sequencing cost and higher computational capabilities have made it possible to sequence complete metagenomes using shotgun metagenomic libraries and assembly of near complete individual genomes from mixed sequences. These efforts are a major advancement from the limited genes identification for biotechnological applications to comprehensive genomic data, which would resolve organism-specific pathways and provide insight into population structure, speciation and evolution. Two almost complete genomes of Leptospirillum group II and Ferroplasma type II, and partial recovery of three other genomes were obtained by shotgun sequencing of DNA from a natural acidophilic biofilm. Sequence analysis revealed pathways of carbon and nitrogen fixation and energy generation for each organism of the simple biofilm community. It also suggested role of Leptospirillum group II in biofilm formation and possibility of Leptospirillum group II and III as first colonizers of the biofilm. Venter et al. sequenced environmental genome of the Sargasso Sea in an attempt to analyse gene contents, diversity and relative abundance of the organisms in the sample. The data obtained by this major sequencing effort led to identification of more than 1.2 million previously unknown genes and assembly of groups of scaffolds closely related to Burkholderia, Shewanella oneidensis, SAR 86 and a conglomerate of Prochlorococcus strains. Phylogenetic analysis of the sequence of 16S rRNA genes revealed presence of 148 previously unknown phylotypes based on 97% similarity cut-off to distinguish unique phylotypes. An ammonium monoxygenase gene was found on an archaeal-associated scaffold indicating their role in oceanic nitrification. Only members of bacterial domain were known for ammonium monoxygenase prior to this study. The presence of genes for transport of phosphonates, utilization of polyphosphates and pyrophosphates, phosphorus transporter \(psrS\) and \(pho\) regulon group of genes indicated capability in the Sargasso Sea’s microbial community to survive in an extremely phosphate-limited environment.

Conclusions

Culture-independent microbial diversity analysis in the last decade has revealed previously uncharacterized members in both bacterial and archaeal domains. These novel ‘yet to be cultured’ bacteria represent an unexplored and unexploited vast gene pool. Attempts to culture these unique bacteria or screening of metagenomic libraries, circumventing the need of culture, have found an impetus in recent times. Development of new culture techniques and more innovative tools in molecular biology related to genomic library construction in culturable members of various bacterial groups are expected to revolutionize the field of biocatalysts and drug discovery. Further decrease in sequencing cost is expected to increase efforts for microbial community genome sequencing in an attempt to understand community structure and ecosystem function. Availability of community genome sequences will help in development of gene expression profiles and physiological studies providing a comprehensive approach to environmental biology.


ACKNOWLEDGEMENTS. We thank Prof. S. K. Brahmachari, Director, IGBR for encouragement and support. R.R. and A.G. acknowledge CSIR for junior research fellowships. R.S. thanks DST (FTP/LS-277/2000) and CSIR (SMM0002) for financial support.
ENVIRONMENTAL GENOMICS:
EXPLORING AND EXPLOITING THE
‘UNCULTURABLE’ MICROBIAL DIVERSITY

Rakesh Sharma*, Ravi Ranjan, Raj Kishor Kapardar, Nar Singh Chauhan and Amit Grover

INTRODUCTION

Men have utilized bacterial products even before they realized their (bacterial) existence. Discovery of antibiotic, antifungal, anticancer, immuno-suppressant compounds and variety of biocatalysts from bacteria in past century encouraged microbiologists to isolate and screen bacteria from diverse habitats. Most of these studies utilized culture-based techniques for isolation of bacteria, assuming culturing reveals most bacteria of an environmental sample. Research in last two decades revealed that this assumption was not correct. Comparison between direct microscopic cell count with colonies growing on the nutrient agar showed that less than one cell in thousand formed a colony from the sample. This means that only about 0.1-1.0 % of the bacteria present in an environmental sample can be cultured by existing laboratory techniques (Kellenberger et al. 2001). Only few thousand cultured bacteria has been described in literature, whereas estimates of bacterial diversity range from 1 million to 100 million (Short, 1997).

The methods, which are independent of culturing are important to characterize the diversity of bacteria in the environment, one such method is the molecular phylogenetic analysis of small subunit ribosomal RNA gene sequence. This molecular phylogenetic technique has revealed that a large number of the uncultured bacteria represent enormous diversity of as yet uncharacterized bacteria (Bintrim et al. 1997, Hugenholtz et al. 1998). This approach has resulted in the discovery of entirely new phylogenetic lineages, some of which are major constituents of the environmental communities that were not detected by traditional cultivation techniques. Results from application of these culture-independent methods to a large number of diverse environments confirmed that our view of bacterial diversity was limited. This points to a wealth of novel and environmentally important bacterial diversity, which is yet to be studied. The molecular phylogenetic analysis of bacterial communities by culture-independent studies resulted in increase of identifiable bacterial divisions to 40. Thirteen of these divisions are characterized by only

*Corresponding author
environmental DNA sequences from "unculturable" bacteria and are so called candidate divisions (Hugenholtz et al. 1998). Similarly, highly clustered novel group of archaea has been identified from marine (Furhman et al. 1992) and terrestrial (Bintrim et al. 1997) environments indicating that archaea are more diverse than previously thought. Whereas culture based techniques has resulted in isolation of archaea only from extreme environmental niches, such as hot springs and thermal vents. Analysis of 16S r-RNA gene clones from soil and marine water samples has helped in identification of organisms having no close association to any cultivated member of archaea (Furhman et al. 1992, Bintrim et al. 1997). The above-mentioned studies show that bacteria in nature are more phylogenetically diverse than has
been accounted for by sequence analysis of cultured strains (Bintrim et al. 1997, Hugenholtz et al. 1998). There is a wealth of novel bacterial types in natural samples that have not been previously described by microbiologists. Presently, much research is aimed at further characterization of these uncultured bacteria, to identify them, to understand their distribution, their role in ecosystem and to devise new culture techniques to cultivate them (Button et al. 1993, Wate et al. 2000, Connan et al. 2002). At the same time these previously unknown "yet to be cultured" bacteria represent a vast and untapped resource for biotechnological exploitation. Molecular biologists in last decade overcome the difficulties in isolation of DNA directly from environmental samples representing genomes of whole microbial community of the sample. This community DNA (metagenome) gave an opportunity to access the gene pool of vast diversity of "unculturable" bacteria circumventing the need to culture. The metagenome form a particular habitat can be exploited for various applications by PCR amplification of genes followed by cloning or by preparation of metagenomic libraries in culturable host. Metagenomic libraries can be screened for functional clones for sequencing and characterization (functional metagenomics) or can be sequenced directly to know more about a particular "unculturable" bacteria or whole microbial community (metagenomics).

**Functional Metagenomics**

Metagenomic DNA from various habitats has been exploited by PCR-based or metagenomic library-based methods in last 5-6 years for discovery of novel biocatalysts or bioactive compounds (Bell et al. 2002, Eschenfeldt et al. 2001, Cottrell et al. 2000). Various commercially important enzymes viz lipase/esterase, amylase, chitinase, xylanase, nitrilase, alcohol oxidase were isolated and sequenced. Some of these biocatalysts were also characterized for their various industrially important properties such as substrate specificity. Many attempts were focused on preparation of large insert libraries to identify novel drug molecules. Some of the important biocatalysts and bioactive compounds identified from metagenomes are briefly described below:

**Biocatalysts**

**Lipase/Esterase:** A lipase gene with less than 20 % similarity to known lipases at amino acid level was cloned from soil DNA using primers against the conserved domains of the known lipases and genome walking to obtain full length gene (Bell et al. 2002). One lipase and three esterase clones were obtained by screening three soil metagenomic libraries by Henne et al. (2002). The deduced protein sequence of these four clones showed less than 50 % amino acid identity to known sequences in the databases. One lipase clone was obtained from a Bacterial Artificial Chromosome (BAC) metagenomic library prepared from soil DNA (Rondon et al. 2000). The clone was not further characterized. In a recent study six lipolytic genes were identified
and partially characterized from a soil library (Lee et al. 2004). These genes encoded lipolytic enzymes, which belonged to hormone sensitive lipase family and showed low homology with the known lipolytic proteins in the databases. The isolation of lipolytic genes with low amino acid homology indicated that lipolytic enzymes represent vast sequence diversity in natural microbial communities and it will not be possible to explore lipolytic genes only by PCR technique from metagenome with the primers designed from the sequence information available from the cultured bacteria in databases.

**Amylase:** An amylase with exceptional process compatibility and economics was obtained by combination of environmental library screening and laboratory evolution (Richardson et al. 2002). It was thermostable, worked at pH 4.5 and did not require calcium ions for activation, which are desirable characteristics of an industrial amylase. In a recent report, an amylase clone was obtained from screening of 30000 clones from a soil metagenomic library (Yun et al. 2004). This enzyme had a temperature optimum of 42ºC and pH optimum of 9.0. This enzyme showed high transglycosylation activity. From the substrate profile the enzyme was regarded as an intermediate type of maltogenic amylase, α-amylase and 4-α-glucanotransferase.

**Chitinase:** Chitinase genes were isolated from coastal and estuarine water metagenomic library using lambda phage based vector. Screening of 230000 clones from costal library resulted in isolation of 432 clones resulting in hydrolysis of 4-methyl-umbelliferyl α-D-N,N'-diacetylchitobioside. Screening of estuarine library resulted in isolation of nine chitinase positive clones (Cotrell et al. 1999). These results were consistent with culture-based estimates of portion of marine bacteria that degrade chitin. PCR method was used to explore the diversity of chitinase genes culturable and "unculturable" marine bacteria (Cottrell et al. 2000).

**Alcohol oxidoreductase:** Combination of traditional enrichment culture and metagenomic approach was used to generate a gene bank of alcohol oxidoreductases containing 24 clones. These clones were completely sequenced. Sixteen of them conferred carbonyl forming phenotype and eight clones exhibited NAD(H) dependent alcohol oxidoreductase activity (Knietsch et al. 2003).

**Nitrilase:** Screening of more than 600 environmental DNA libraries containing $10^6$ to $10^{10}$ members per library by selection based expression assay resulted in isolation of 137 nitrilases. Nitrilase genes were targeted in a selection based expression assay of clonal population numbering $10^6$ to $10^{10}$ members per e-DNA library. Sequence analysis revealed five major sequence clades within the nitrilase subfamily. They characterize substrate specificity & stereo selectivity by using three-nitrile substrate to use in chiral pharmaceutical synthesis. (Robertson et al. 2004).

**Xylanase:** Libraries prepared from the intestinal tracts of termites and moths were screened to obtain unusual microbial xylanases. Four unique clones that generated clearing zone on azo linked xylan agar plate were isolated. Biochemical analysis revealed that they were true xylanases though their sequences were remarkably different from the known xylanases (Brennan et al. 2004)
Diketo-D-Gluonic acid reductase: PCR amplification and genome walking techniques were used for isolation of two complete genes for 2,5-diketo-D-gluconic acid reductase. Compared to previously known 2,5-diketo-D-gluconic acid reductases, they had some valuable properties like lower $K_m$ values and higher thermostability. These environmental DKGRS accepted NADH and NADPH as co-substrate (Eschenfeldt et al. 2001).

Hydroxybutyrate dehydrogenase: Three different soil libraries containing more than 9,30,000 recombinants were constructed and screened for various enzymatic activities. Screening resulted in isolation of five clones that grow with 4-hydroxybutyrate as sole carbon and energy source. Gene sequences from three clones showed similarity to known genes of 4-hydroxybutyrate dehydrogenase, member of DedA family and enoyl coenzymeA hydratase/isomerase. The other two sequenced inserts showed no homology to known genes in databases (Henne et al. 1999).

Other enzymes: More than 15 different genes encoding biocatalysts, including a stereoselective amidase, two cellulases, a 1,4-á-glucan branching enzyme, amylase and two pectate lyases were identified from a mixed microbial population (Voget et al. 2003). In a recent report, six amidase genes were identified from metagenomic libraries. One of the amidase exhibited high activity toward penicillin (Gabor et al. 2004).

Bioactive compounds

Long chain acyl-tyrosine antibiotics: Isolation of long chain acyl-tyrosine antibiotics, derived from a single environmental DNA open reading frame, were reported in another study (Brady et al. 2000), followed by identification of a biosynthetic gene cluster that produced two additional families of natural products derived from long chain N-acyl-tyrosines (Brady et al. 2002).

Violacein: A blue clone was isolated and characterized from environmental DNA cosmid library. Colour producing clones often indicate presence of small molecules. Characterization of blue clone revealed presence of a four-gene biosynthetic cluster conferring production of violacein and deoxyviolacein in E. coli host. The DNA sequence of the gene cluster resembled with the violacein gene cluster sequenced from cultured bacterium Chromobacterium violaceum (Brady et al. 2001).

Turbomycin A and B: Characterization of dark brown color producing BAC metagenomic clones revealed presence of broad spectrum antibiotic compounds Turbomycin A and Turbomycin B. Sequence analysis of the clones revealed that a single open reading frame was necessary and sufficient to confer brown, orange and red pigments on E. coli. It was demonstrated that interaction of indole with homogentististic acid, synthesized by 4-hydroxy phenyl pyruvate dioxygenase encoded by open reading frame from the BAC, resulted in production of Turbomycin A and Turbomycin B in recombinant clone (Gillispie et al. 2002).

Indirubin: A BAC soil DNA library, containing inserts between 5-120 Kb resulted in isolation of many antibacterial clones. One of these clones contained indirubin and related small molecules (Macneil et al. 2001).
**Metagenomics**

Culture-independent microbial diversity analysis methods revealed presence of vast diversity of yet uncharacterized bacteria but most of these methods fail to indicate about the physiology of these "unculturable" bacteria and their role in ecosystem. Genomic information of the "unculturable" bacteria can be obtained from the clones in metagenomic libraries, which can help in understanding their physiology and also about their role in ecosystem. Some success has been obtained in achieving these goals in recent years, which include sequence snapshots of rRNA genes containing BAC clones to large scale sequencing of metagenomes of acidophilic biofilms and Sargasso Sea.

**Bacterial rhodopsins:** Sequence analysis of a 130 Kb genomic fragment from a BAC clone that encoded the r-RNA operon from an uncultivated member of marine g-proteobacteria (SAR 86 group) revealed presence of rhodopsin gene. This gene encoded a protein with highest amino acid sequence similarity with archaeal rhodopsins. The halophilic archaea contain bacteriorhodopsins that function as light-driven proton pump for energy generation. This first report of presence of rhodopsin in marine g-proteobacteria in domain bacteria suggested the possibility of a previously unrecognized phototrophic pathway that may influence the flux of carbon and energy in the ocean's photozone worldwide (Beza et al. 2000). Rhodopsins are now known to occur both in α- and β-proteobacterial lineages. Sequence comparison on BAC clones carrying 16S rRNA genes and bacteriorhodopsin genes from different SAR86 subgroups revealed different rhodopsin sequence types within same SAR86 rRNA subgroup (Sabeehi et al. 2004).

**Creanoarchaeta:** The DNA polymerase gene of *Crenarchaeum symbiosum* was identified from vicinity of the rRNA operon on a large genomic fragment. Encoded protein showed highest similarity with archaeal family B (α-type) DNA polymerases (Schleper et al. 1997). Characterization of this enzymes revealed that it was a thermolabile enzyme contrary to its close homologues from other archaea. This was a first report of characterization of a protein from nonthermophilic uncultured crenarchaeote. The contents and structural comparison of 34 Kb fragment containing 16S/23S rRNA operon and 17 genes encoding putative proteins from a non-thermophilic clade of soil crenarchaeota (archaea) revealed significant differences from their previously studied marine relative (Quaiser et al. 2002).

**Acidobacteria:** In order to understand physiology of the yet to be cultured members of the acidobacteria division, a BAC clone containing full r-RNA gene and 20 other open reading frames were completely sequenced (Liles et al. 2003). Genes for cell division, cell cycling, folic acid biosynthesis, substrate metabolism, amino acid uptake, DNA repair and transcriptional regulation were identified.

**Acid mine drainage community:** In a recent and first report of its kind, Tyson et al (2004) reported sequencing of metagenome of a biofilm bacterial community of acid mine drainage system in California. The metagenomic library prepared from the community representing at least six major phylotypes, with three bacterial and
three archaeal lineages, was used for sequencing by shotgun approach. End sequencing of about 50,000 clones helped in assembly of two near complete genomes and partial recovery of three other genomes. Genome sequence analysis for each member revealed genes for various metabolic pathways utilized for carbon and nitrogen fixation and energy generation. Number of genes identified for proton efflux systems, antiporters, symporters and genes for resistance to copper, arsenite, mercury, zinc, silver and cadmium indicated towards the survival strategies utilized by the members of this biofilm community in the adverse environmental conditions they live.

**Sargossa Sea Microbial Community:** In another path breaking recent report Venter *et al.* (2004) reported shotgun sequencing of 1.045 billion base pairs from metagenomic libraries prepared from Sargasso Sea water communities. This sequence data resulted in assembly of distinct groups of scaffolds belonging to strain related to *Burkholderia*, strains closely related *Shewanella oneidensis*, SAR86, *Prochlorococcus* strains and an uncultured archaean. Ten mega plasmid sequences were also identified. More than 1.2 million previously unknown genes were identified including 782 new rhodopsin-like photoreceptors. This data also provided insights in to nitrogen fixation and phosphorus utilization in Sargasso Sea microbial communities.

**Conclusions**

Large proportion of bacterial diversity remains inaccessible to microbiologists because of its “unculturable” status. It is important to understand these “yet to cultured bacteria” by culture-independent approaches to get insight in their richness, abundance and ecological significance. Phylogenetic and genomic information about this bacterial majority is helping in devising newer cultivation strategies and many members of previously “unculturable” groups have been successfully obtained in culture in recent years. Metagenomic approach is allowing modern biologists to access gene pool of whole microbial communities without the culture bias. This has opened up a great vista for novel biocatalysts and bioactive compounds discovery. Shotgun sequencing of whole microbial communities along with microarray profiling of environmental samples is expected to boost our current understanding of microbial ecology and ecosystem functioning.

**Acknowledgements**

We thank Prof. S.K. Brahmachari, Director, I.G.I.B. for encouragement and support. AG, RR and NSC acknowledge C.S.I.R. for Junior Research Fellowship. RS acknowledge financial support from CSIR (project # SMM0002) and DST (project # SR/FTP/LS-277/2000).
REFERENCES


