NOTE

Susan Grant · Dimitry Y. Sorokin · William D. Grant
Brian E. Jones · Shaun Heaphy

A phylogenetic analysis of Wadi el Natrun soda lake cellulase enrichment cultures and identification of cellulase genes from these cultures

Abstract Samples of sediments and surrounding soda soils (SS) from the extremely saline and alkaline lakes of the Wadi el Natrun in the Libyan Desert, Egypt, were obtained in October 2000. Anaerobic enrichment cultures were grown from these samples, DNA isolated, and the bacterial diversity assessed by 16S rRNA gene clone analysis. Clones derived from lake sediments (LS) most closely matched *Clostridium* spp., *Natronoincola histidinovorans*, *Halocella cellulolytica*, *Bacillus* spp., and the *Cytophaga–Flexibacter–Bacteroides* group. Similar clones were identified in the SS, but *Bacillus* spp. predominated. Many of the clones were most closely related to organisms already identified in alkaline or saline environments. Two genomic DNA libraries were made from the pooled LS enrichments and a single SS-enrichment sample. A novel cellulase activity was identified and characterized in each. The lake cellulase ORF encoded a protein of 1,118 amino acids; BLASTP analysis showed it was most closely related to an endoglucanase from *Xanthomonas campestris*. The soil-cellulase ORF encoded a protein of 634 amino acids that was most closely related to an endoglucanase from *Fibrobacter succinogenes*.

S. Grant (✉) · W. D. Grant · S. Heaphy
Department of Microbiology and Immunology,
University of Leicester, University Road,
Leicester, LE1 9HN, UK
E-mail: sg23@le.ac.uk
Tel.: +44-116-2522940
Fax: +44-116-2525030

D. Y. Sorokin
Department of Biotechnology,
Delft University of Technology, Delft, The Netherlands

D. Y. Sorokin
Institute of Microbiology, Russian Academy of Science,
17811 Moscow, Russia

B. E. Jones
Genencor International B, V., Archimedesweg 30,
2333 CN Leiden, The Netherlands

Soda lakes, with high salinity and alkalinity, are extreme but often highly productive environments (Jones et al. 1998). Studies, particularly on the lakes of the East African Rift Valley in Kenya, have shown that they are habitats for novel species of bacteria and archaea (Duckworth et al. 1996). Soda lakes are found all over the world, but how the prokaryotic species distribution varies between geographically disperse sites is largely unknown. The extremophilic nature of these organisms makes them a potential source of enzymes and metabolites with industrial uses (Horikoshi 1999; Marrs et al. 1999; Margesin and Schinner 2001). However, in common with microbes in other habitats, most have not, and cannot be, cultured in the laboratory (Amann et al. 1995; Grant et al. 1999). A now standard approach to enable access to novel enzymes from the total microbial population, including the unculturables, is to construct genomic DNA libraries, using material extracted directly from environmental samples (Henne et al. 2000; Rees et al. 2003a). Similarly, 16S rRNA sequence analysis of the microbial DNA allows determination of the microbial diversity.

A wide variety of enzyme classes have been isolated from cloned environmental DNA including hydrolases, but this approach is not always successful. Some environmental samples contain very little DNA; in others, the DNA can be very difficult to extract and clone; in yet others, the desired enzyme activity cannot be detected. In these circumstances it may be desirable to use enrichment cultures selecting for bacteria likely to have the desired enzyme activity. The biodiversity of an enrichment culture will not be as rich as the original environmental source because of the nonculturability of many organisms (Rees et al. 2003b). Nonetheless, T-RFLP analysis of three different enrichment cultures derived from soda-lake samples detected between 95 and 107 prokaryotic species (Rees et al. 2003b). There is also evidence suggesting that specific enzyme activities can be

isolated from libraries prepared from genomic DNA isolated from enrichment cultures, with a higher frequency than from the environmental DNA itself. Rees et al. (2003a) found that the incidence of esterase- and cellulase-gene recovery from libraries made from the DNA isolated from enrichment cultures grown, respectively, with olive oil and cellulose as the major carbon sources is increased three- to fourfold compared to libraries made directly from the total environmental DNA.

Genomic DNA libraries made from enrichment cultures that select for particular enzyme activities can therefore be a valuable resource. Accordingly, we constructed two libraries, one from pooled anaerobic, cellulose-enrichment cultures of Wadi el Natrun lake sediments (LS) and another from a Wadi el Natrun soda soil (SS) enrichment, in alkaline medium containing microcrystalline cellulose. These libraries were then screened for endocellulase activity. In addition, the biodiversity in these enrichments was assessed following 16S rDNA amplification of the genomic DNA and sequence analysis of a number of clones. This is the first such phylogenetic study of this particular environment.

Samples were collected from the Wadi el Natrun soda lakes in Egypt in October 2000 (Table 1). The basic hydrochemical and microbiological characteristics of these lakes have been described by Taher (1999) and Imhoff et al. (1979), respectively. At the time of sampling, the water pH and total salt content of the lakes were within the ranges 9.1–10.4 and 201–360 g l^{-1}, respectively. The dominant cation in the lakes is Na^+

(2.1–4.5 M), with traces of Mg^{2+}, K^+, and Ca^{2+} (each less than 0.05 M in any of the lakes). The main anion was Cl^- (2.1–4.5 M), with lesser amounts of HCO_3^-, CO_3^{2-}, and SO_4^{2-}. The carbonate/bicarbonate alkalinity varied from 0.1–0.8 M, and sulfate was less than 0.3 M. With concentrations of cations other than sodium being very low, virtually all the chloride is presumed to be present as sodium chloride. The molarities of chloride ions shown in Table 1 therefore reflect the sodium chloride concentrations.

A total of eight sediment cores (the top 10 cm) were collected from eight different lakes, and three samples of the top 5 cm of sandy soil covered with soda crust were obtained near three of the lakes. Samples were stored at 4°C for 5 days before use.

All these LS and SS samples were tested for cellulose degradation. The liquid medium for the enrichments contained (g l^{-1}) Na_2CO_3 (22.0), $NaHCO_3$ (8.0), NaCl (5.0), K_2HPO_4 (1.0), NH_4Cl (0.1), microcrystalline cellulose (Whatman, Maidstone, UK) (5.0), $MgCl_2.6H_2O$ (0.1), plus 1 ml trace metals solution (Pfennig and Lippert 1966), pH 10.0. Cellulose degradation was found to increase by utilizing this relatively low sodium chloride concentration.

Only one SS sample showed visible cellulose degradation. This sample was of sandy soil with a soda crust taken from near Lake Khadra. The pH of a 1:5 water extract was 10.0. After prolonged incubation (3–4 weeks at 30°C) under static (microaerobic) conditions, the culture appeared yellowish and showed visible cellulose degradation. This increased after transfer to anaerobic

Table 1 Samples from the Wadi el Natrun (October 2000)

Lake	Sediment pH (diluted 1:5)	CO_3^{2-} (M)	Cl^- (M)	Total dissolved salts (g l^{-1})	Description of the lakes	Description of soil sample
Hamra	10.1	0.23	Not tested	201	Full of brine, shallow, little or no salts around, water colorless, mats	
Umm-Risha	9.9	0.16	4.54	320	Almost dry, huge salt deposition, salterns, brine is pink, thick salt crust covers the sediments	Sand with soda crust, pH of water extract (1:5) 10.3
Fazda	10.1	0.46	2.11	250	Half-dry, salt depositions, salterns, brine is green-yellowish, anaerobic with sulfide smell, hard salt crust covers the sediments, mats of cyanobacteria	
Beidah	9.8	0.28	4.50	320	Full of brine, shallow, thick salt depositions, brine almost colorless	Sandy clay, pH of water extract (1:5) 10.1
Gaara	9.8	0.21	3.21	260	Full of brine, shallow, salt crust on the sediment surface, brine is orange	
Khadra	10.3	0.67	Not tested	360	Small, shallow, very thick salt crust covers sediments, brine is red and contains a lot of pink biomass, mats of cyanobacteria	Sand with soda crust, pH of water extract (1:5) 10.0
Zugm	10.2	0.75	3.94	360	Almost dry, salt depositions around and on the sediment surface, brine is red	
Ruzita	9.7	0.11	4.06	340	Half-dry, shallow, salt depositions and saltern, thick mats of cyanobacteria, brine is deep pink	

conditions and by addition of 10 mM Na_2SO_4 to the above medium. Without sulfate addition, cellulose degradation was very weak. Sulfide was accumulated in millimole concentrations during cellulose degradation. The liquid phase developed a dense monoculture of large motile rods, while the cellulose sediment was dominated by thin, vibroid cells attached to the cellulose crystals. Attempts to culture these bacterial species on solid medium were unsuccessful. The liquid and cellulose fractions were therefore pelleted from 10 ml culture and stored frozen at −80°C. DNA extracted from these two fractions was later pooled for construction of the SS library.

Three LS samples also proved positive for degradation of microcrystalline cellulose when enrichments were grown under anaerobic conditions with the addition of sulfate. Without the addition of sulfate, no visible cellulose degradation occurred. These samples were from lakes Fazda, Gaara, and Ruzita, pHs of 10.1, 9.8, and 9.7, respectively, for a 1:5 dilution, and comprised black, anaerobic mud rich in organic carbon and sulfide. Liquid anaerobic enrichments gave visible growth after 3 weeks of incubation with each of the three samples. The color of the Fazda culture turned yellowish, while those from the Ruzita and Gaara samples were reddish. Microscopic examination showed two dominant morphotypes. In the Fazda culture, coccoid cells forming microcolonies firmly attached to the cellulose crystals were clearly visible. In solution, thin filaments with a round spore at one end were dominant. The Ruzita culture showed a similar population, but the coccoid component was not so abundant.

Accumulation of 5–8 mM of sulfide was detected in the Fazda and Ruzita samples, but not in the Gaara enrichment. At this stage, cells from 5–10 ml of the three cultures were pelleted and stored frozen at −80°C for DNA extraction. The Fazda and Ruzita cultures were subcultured to fresh sulfate-containing media. After incubation, both cultures turned yellowish, with a dominant population of vibroid cells attached to cellulose (similar to the SS enrichment culture from Lake Khadra), and rapid anaerobic cellulose degradation was observed in both. Material from these transfers was also pelleted for DNA extraction. The second LS library was constructed after pooling the DNA from all of these samples.

DNA was extracted from the pelleted cell mass, using the Pharmacia Biotech GenomicPrep Cells and Tissue DNA Isolation kit, following the manufacturer's protocol, except that incubation with 50 μl lysozyme solution (50 mg ml^{-1} in 10 mM Tris-HCl pH 8.0, 1 mM EDTA) for 30 min at 37°C preceded the addition of the lysis solution provided in the kit in order to facilitate degradation of Gram-positive cell walls. The DNA samples were used to construct genomic λ DNA libraries, pooling the samples from the LS enrichments for one library and the SS enrichments for the second. Partial restriction digestion with *Sau*3A1 was carried out to give fragments concentrated in the 2–10 kb size range.

Restricted DNA was run out on 0.5% TAE agarose gels, using Pharmacia N/A Agarose. Fragments of 2–10 kb were excised and concentrated by reversed-current electrophoresis, and then extracted from the agarose gel, using the QIAEX II system (QIAGEN), according to the manufacturer's protocol. The restricted DNA fragments were ligated into the ZAP Express vector (predigested with *Bam*H1 and phosphatase treated), and then packaged with the Gigapack III Gold packaging kit (Stratagene). The primary libraries were titered with *Escherichia coli* XLI blue MRF'. Blue–white screening in the presence of Xgal and IPTG was used to determine the cloning efficiency. Libraries were amplified and stored at −80°C, following the manufacturer's instructions. Library quality was also checked by assessing the size of the inserts in 24 of the clones. Phagemid pBKCMV was excised from the ZAP Express libraries, using ExAssist helper phage, and used to infect *E. coli* strain XLOLR, according to the Stratagene protocol. Plasmid-containing clones were recovered after plating on LB containing kanamycin (50 μg ml^{-1}). Plasmid DNA was extracted using the Wizard Plus SV Minipreps DNA Purification System (Promega). The insert size was determined by restriction digestion with *Hin*dIII or *Hin*dIII plus *Pst*I, followed by agarose gel electrophoresis.

Detection of cellulolytic activity was based on the method of Teather and Wood (1982). *E. coli* clones containing the pBKCMV plasmids were screened by plating the phagemid-infected XLOLR strain to give approximately 500–750 colonies per 7 cm^2 plate onto LB medium at pH 7.5, containing 0.5% carboxymethylcellulose as substrate and 50 μg ml^{-1} kanamycin. IPTG (15 μl of 0.5 M) was spread on the surface of the agar. After overnight growth, the colonies were overlaid with 2.5 ml 0.7% molten water agarose at 45°C. After this had set, the plates were flooded with an aqueous solution of 1 mg ml^{-1} Congo Red and incubated at room temperature for 30 min with slow rocking. They were then washed twice with 1 M NaCl for 30 min. Positive clones were visible by a yellow zone of hydrolysis against a red background.

Plasmid DNA was isolated from positive clones, and the insert size determined as described above. DNA sequencing was carried out by commercial services. Complete coverage of the sequence was obtained by primer walking from both the 5′ and 3′ ends. Sequences were assembled and compared using programs in the GCG Wisconsin Package, version 10.2-UNIX, available at the University of Leicester. Comparison of sequences to those in the databases was made using BLASTX 2.1.3 on the National Center for Biotechnology Information (NCBI) Web site (http://www.ncbi.nlm.nih.gov; Altschul et al. 1997). ORF Finder at NCBI was used to identify possible open reading frames.

The 16S rDNA was amplified from the same two pooled DNA samples used to make the SS and LS libraries with the primers 27Fb (5′-AGA GTT TGA TCC TGG CTC AG-3′) for bacteria, or 27Fa (5′-TC(CT)

Fig. 1 Alignment of the amino acid sequence of the putative ORF for the Wadi el Natrun lake sediment enrichment cellulose (*top line*) to that of *Xanthomonas campestris* (accession AJ245855, *bottom line*). The *middle line* indicates identity (*capitals*), similarity (*), or nonidentity

```
 13 AAILLSAAWTSAEAYRPFKVNQAGYHPDADKLAVVPVNGSVLTTSQYQIVDVESGDTVFT  72
    AA     A T+  A       VNQ GY P + KLAVV +  S          V+   G  V
 21 AADTTGTAGTTPAAGATIHVNQLGYLPGSAKLAVVGLPASAAAGPDRFTVEDAKGARVLE  80

 73 GALLGGDGLYWEHSDETVLHALFTDLSTPGTYRIEHPDLGESHPFEISEEVLTEVSRSAL  132
    G+L        W + +    A F+ L TPGTYR++   +  S  F I+      V  ++L
 81 GSL--SPAASWAPAGQQARVADFSALRTPGTYRLKIAGMPASDAFPIAASAYQPVVDASL  138

133 KAYYFNRASTEITEPYAGPWARPMGHPDDNAMIHSSAATEHRPVGYTFSSSKGWYDAGDF  192
    KA+Y+NRASTE+   +AG +AR  GHPD   IH SAA+  RP     S+ KGWYDAGD+
139 KAFYYNRASTELPAQFAGRYARAAGHPDTEVRIHPSAASATRPANSVISAPKGWYDAGDY  198

193 NKYIVNSGITTYTLLAAYEHFPNYFEELTVNIPESGSGVPDILDEIRWNLDWMITMQDPH  252
    NKYIVNSGI+TYT+LAAYEH+P  F+   + IP   GVP IL E+ WNL+WM+ MQDP
199 NKYIVNSGISTYTVLAAYEHYPALFKAQALTIPNDAPGVPGILQEVWWNLEWMLAMQDPA  258

253 DGGVYHKLTSPQFSGIIMPHQDNSARYAVQKSTAATLNFAGVMAVASRVYRPYDRDF---  309
    DGGVYHKLT  QF G++MP Q    RY V K+TAATL+FA VMA ASRVY P+D+ +
259 DGGVYHKLTDKQFDGLVMPDQAKQQRYVVMKATAATLDFAAVMAAASRVYAPFDKQYPGA  318

310 SERALAAAEFAWQWALEHPNLYYRQDAMNQEHSPDIVTGEYGDSNVSDEFDWAAAELYIT  369
    S R L AA  AW WA +HPN+ YRQ        D+ TG Y D+ V DEF WAAAELY++
319 SARMLKAARSAWAWAQQHPNVIYRQ-------PDDVRTGGYDDATVDDEFAWAAAELYVS  371

370 TGDDSYWHARNLGNVHGITIPAWPQVRPLAWVSLAHHRDRLTPAANVGHITGRIEQMANE  429
    T +D+++ A    NVH  ++P W  V  LAW+SLA HRD+LTP A+   I G I  +A
372 TREDAFYDAMLARNVHA-SVPGWSNVGGLAWMSLAAHRDQLTPHADRARIEGEISGLAQR  430

430 LLNAHNNSAYLMSMGYRGGGDFVWGSNSVALNHSLMLLQGYRLTGNGSMLDAAQANLDYV  489
    L +     S + ++M      DFVWGSN+V LN ++MLLQGYR+T    LDAAQ+ LDY+
431 LGDQWQASPWRLAM---ADNDFVWGSNAVVLNQAMMLLQGYRVTQQRRYLDAAQSQLDYI  487

490 LGRNPTGYSFVTGIGSRTPMDPHHRPSAAHSHDAPVPGLVVGGPHSGQQDGCN----YPS  545
    LGRNP G SFVTG+G RTPM  HHRPS A   DAPVPG + GGP  GQQD  +     YPS
488 LGRNPLGLSFVTGVGQRTPMHIHHRPSEADGIDAPVPGWLAGGPQPGQQDAKDCKVAYPS  547

546 DLPALSYVDDWCSYATNEVAINWNAPLVYVAGAIDA  581
     L ALSY+D+ CSYATNEVAINWNAPLVYV+ AI+A
548 KLAALSYLDNTCSYATNEVAINWNAPLVYVSAAIEA  583
```

GGT TGA TCC TG(CG) CGG-3′) for archaea, both with the universal reverse primer 1492R (5′-ACG G(ATC)T ACC TTG TTA CGA CTT-3′). PCR primers were obtained from TAGN (Newcastle, UK) and *Taq* polymerase was obtained from ABgene (Epsom, UK). Only the bacterial primers gave a clear PCR product for the two DNA samples. These were cleaned by passing through a QIAquick PCR Purification kit (QIAGEN) and cloned into the pGEM-T Easy vector (Promega). Partial sequences of over 45 clones were obtained for each of the two enrichments, LS and SS. The BLASTN program at NCBI and the FASTA program of the GCG Wisconsin Package were used to identify bacterial species with 16S rDNA sequences closest to those in the clones.

Sequences showing less than 99.5% identity to any other clones were put into Pearson FASTA format and aligned using the ClustalX program, version 1.81 (http://inn-prot.weizmann.ac.il/software/ClustalX.html). Here, they were aligned with sequences from closely related organisms and also that of *E. coli*, which acts as a root in the subsequent tree drawing. The alignment was then transferred into the GeneDoc program (http://www.psc.edu/biomed/genedoc) for final manual editing.

The phylogenetic trees were constructed from the alignments using the TREECON for Windows package (Van de Peer and De Wachter 1994). Distance estimations were calculated using the substitution rate calibration of Jukes and Cantor (1969). Tree topology was inferred by the neighbor-joining method, and the tree was rooted to the *E. coli* sequence.

λ Libraries were made using the DNA described above. The Wadi el Natrun LS anaerobic enrichment library and SS library had primary titers of 4.9×10^5 and 4.5×10^5 clones, respectively. Blue–white screening indicated that over 93% of the clones contained an insert; *Hin*dIII digestion of 24 clones from each library showed that insert sizes varied from 2.0 to 5.5 kb. Screening 35,000 clones from the LS library for endocellulase producers identified four clones containing inserts of 3.5–4.5 kb in size. Screening 37,000 clones from the SS library for endocellulase producers identified 12 clones containing inserts of 2.0–6.0 kb.

The full sequences of the inserts in all 16 clones were obtained by primer walking.

For the four LS clones, the nucleotide sequences were >99.5% identical over common regions, indicating that a single gene had been cloned. The largest insert, LScel4, was 4,513 bp in length and contained a putative cellulase ORF from nucleotides 324 (ATG start codon) to 3,680 (TAG stop), encoding 1,118 amino acids. The complete sequence of the insert has been deposited in the EMBL database (accession number AJ622824). The other three related clones contained smaller inserts with missing fragments off the 3′ end compared to the complete ORF, such that no stop codon was reached. No BLASTN similarities were found with these sequences, but a BLASTP search identified similarities to various bacterial endo-1,4-β-D-glucanase sequences. The highest score was found to a glucanase (accession AJ245855, protein ID CAB63115.1) from *Xanthomonas campestris* (ATCC

33913). There was 51% identity and 63% positive match over amino acids 13–581 from the N-terminal region of the LS cellulase to almost the entire *X. campestris* cellulase, which is 586 amino acids in length. The alignment of these two amino acid sequences to each other is shown in Fig. 1. Almost as significant was the match to a smaller glucanase of 506 amino acids from *Pseudomonas* sp. SK38 506 (accession AF296443), showing 54% identity and 67% positive match over amino acids 83–581 of the LS cellulase against the complete sequence of the *Pseudomonas* protein. The C-terminal part of the LS cellulase from amino acid 600–1,118 showed a much lower identity of 32–35% over about 150 amino acids to putative secreted proteins from *Streptomyces coelicolor* and *Microbulbifer degradans*. It also had low identity (less than 30%) to hydrolases such as chitinases, mannanases, cellulases, and xylanases from a variety of species.

Fig. 2 Alignment of the amino acid sequence of the putative ORF for the Wadi el Natrun soda soil enrichment cellulase (*top line*) to the endoglucanase F precursor from *F. succinogenes* (accession U39070) (*bottom line*). The *middle line* indicates identity (*capitals*), similarity (***), or nonidentity

```
  2   NENRVGDGNPVILYVYDVKFIEEIPGYVCPEEFWGAFESDAPDLQLHDFATEQDRN-WET  60
      N   G   PV ++V  + F   I  +V P+   W +F+S+APD   + DF  +  ++ WE
403   NAGEPGKPAPVTVFVDQITFTSNID-WVDPDLKWDSFKSNAPDYVISDFEGKFAKDKWEP  461

 61   STGPKSSGSFEFVEAAPEESEGYSIKISYSLHDWVDIIYNYRDNNSPAEHRDWTNHWGLK  120
      STGPKS   F+    A  +S   +I+  Y L DWVD++ + + N  PA  RDWT HWG+
462   STGPKSQLKFKVENCAEFKSNCLNIE-HYLLADWVDVVLDMQKNGRPAADRDWTKHWGIM  520

121   FDVYTERAFQPFNVQVNDSGNELWIATAGAEKGWTEVIVAFRDFYKFPHWQPPDAVHTGN  180
      FDVY+E+A+Q   VQV D+GNE++++  GA KG T ++V FR F KFP++QPP+AV  G
521   FDVYSEKAWQSITVQVQDAGNEIFVSNVGAPKGKTTILVPFRTFGKFPYYQPPNAVENGL  580

181   FDIENVIALDFKPSGEGTSAQFKVHNVRLTNDREARRAPVPEYVAVNVTGSFDEVVTEQI  240
      FD++ V ALDFKPSGEGT+   FKV N+RLTN RE +    P  + V VG D V+   I
581   FDLKGVTALDFKPSGEGTAGGFKVDNIRLTNQREVKAKERPAVIKVLVKGEKD-VLNPNI  639

241   NEGIFGINAALWDGDLL-----LPETVEYVKAVNHHVLRYPGGLRADDDNWKEVLAAQDW  295
      + G+FGINAALWDGD+L        +T EY K +NH ++RYPGGLRADDD+WKE+L   DW
640   SGGLFGINAALWDGDMLDNKNFKVQTREYAKRINHGIIRYPGGLRADDDHWKEILDNHDW  699

296   MVDTDQFLEFCRATNTEAMITVNFGSGTVEDAADWVRHHNVENDDNVRYWEVGNELYGDW  355
      MVDTD+FLE+ + T + AM TVNFGSGTV++AADWV H N++       + YWE+GNE+YG+W
700   MVDTDEFLEWLKKTGSNAMFTVNFGSGTVKEAADWVKHTNIDKKAGILYWEIGNEIYGNW  759

356   HP-FQTTAED----YGKRTREFIIAMKEVDPTIKVTFVGVLDGEWNRIVLEHVKDVADGI  410
      HP ++   +D     YGKR R+FI AMK+VDPTIKV  +GVL+G+WN  VL    D+ADG+
760   HPYYEKYGKDGGTIYGKRARKFIEAMKKVDPTIKVAVLGVLEGDWNDKVLAETGDIADGL  819

411   NVHHYPQTTGEENDAGLLSSPQTLDDIIPSVRRQLAEWGE---PGREYEIWLTEWNSVDF  467
       VHHYPQ  GEEND +LS+PQTL  I   + + + +W +      ++ E+WLTEWNSVDF
820   IVHHYPQHFGEENDFAMLSAPQTLTAIYERLHKVVDKWTKKYNKDKKIELWLTEWNSVDF  879

468   EPGPQTLSIVNGLFVADYLGMLAKHNIEQASYWDIHNDITDQGGDYGYLSRTGAPDGCNV  527
       PGPQTLS+ NGLFVADYLGMLA  N++ A YWDIHNDIT +GGDYGYL+R+G  +  N
880   NPGPQTLSVENGLFVADYLGMLATENVDNAQYWDIHNDITPEGGDYGYLTRSG-EECMNC  938

528   PRPSYWAFKMASHSLGRGSLLNS---GTESDYVTSYLTYDNGKKSLMLVNKYPQTRADVT  584
      PRPSYWAF+MAS +L RG L+ +     G E   +T+Y T +  KK L+LVNK P +  D+
939   PRPSYWAFQMASDAL-RGKLMKTTIKGDEDALLTAYWTVNGNKKQLLLVNKSPYSEFDIK  997

585   INIPGFEGRGTKQVL--TAESGQPG----PEKGNFRVRQGTELALPPYSITTITLD  634
      ++IPGF+G+    Q L  T+E + G     P K   V       + +  ++T ITL+
998   LDIPGFKGKAKVQTLDKTSEKLKEGWANDPSKKAKTVDISKGIKVGKRTLTLITLE  1053
```

Fig. 3 Rooted phylogenetic tree showing the relationship of randomly selected bacterial 16S rDNA clones from the Wadi el Natrun lake sediment enrichments to each other and to related sequences in the databases, of both cultured organisms and uncultured clones. The tree was rooted with the *Escherichia coli* sequence. The *scale bar* represents the number of inferred nucleotide substitutions per site. *Values at nodes* indicate >50% percentage occurrence in 1,000 bootstrapped trees


0.1 substitutions/site
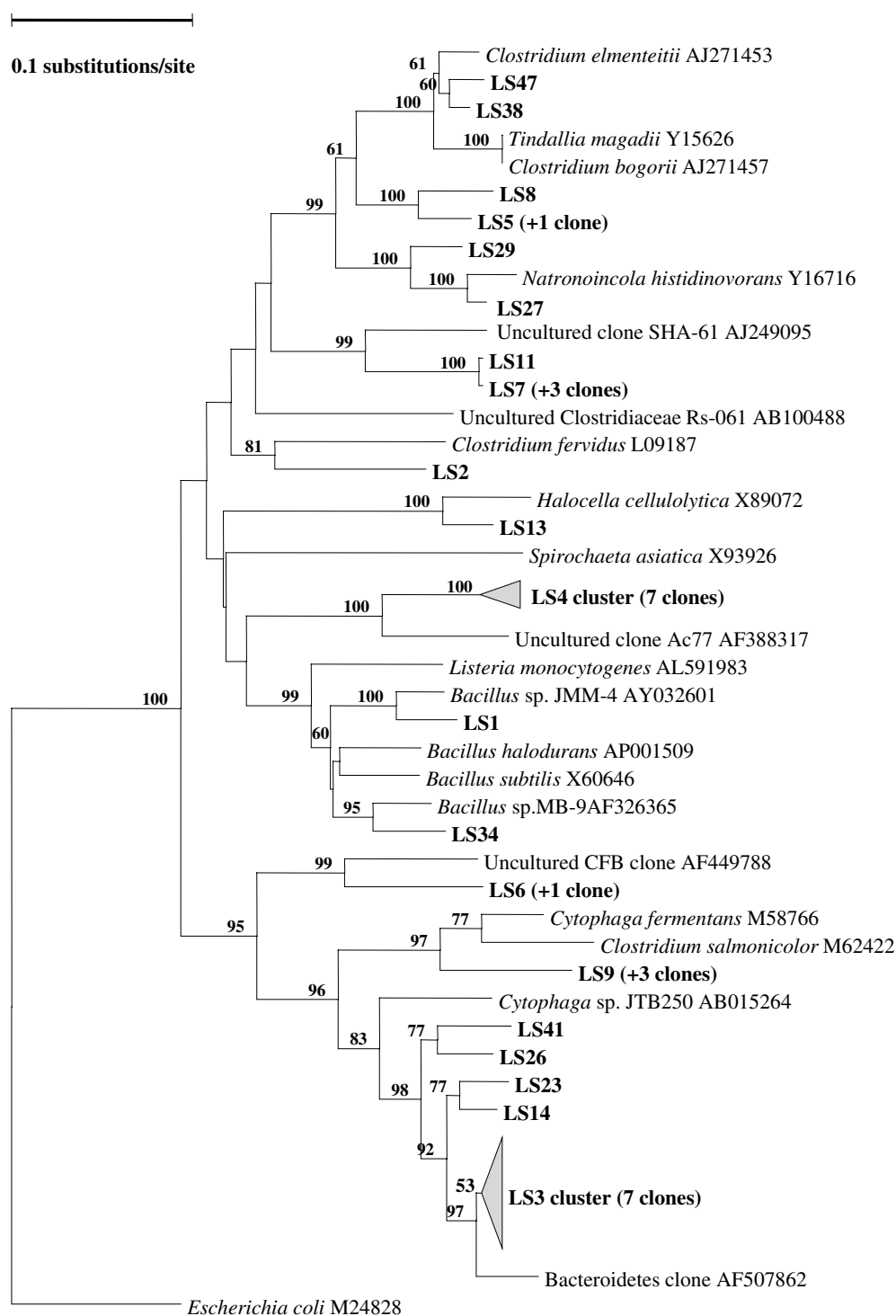
A protein-family conserved domain search, using the ORF cognitor function at NCBI, showed three significant alignments. Amino acids 125–578 aligned to the glycosyl-hydrolase family 9 conserved domain (pfam 00759); residues 29–115 aligned to an immunoglobulin-like domain (pfam 02927) found in a number of cellulases; and amino acids 678–795 showed significant alignment to an F5/F8 type C (discoidin) domain (pfam 00754), which is found on membrane-associated proteins and may be a carbohydrate-binding domain. The LS cellulase may therefore be a cell-surface cellulase or a secreted protein.

Other putative ORFs present in the cloned insert were much smaller (217 residues or less) and had low identities to mostly hypothetical proteins.

Screening of the SS library resulted in 12 positive cellulase clones with six different insert sizes ranging from 1,911–5,773 bp, suggesting that six independently

Fig. 4 Rooted phylogenetic tree showing the relationship of randomly selected bacterial 16S rDNA clones from the Wadi el Natrun soda soil enrichment to each other and to related sequences in the databases, of both cultured organisms and uncultured clones. The tree was rooted with the *E. coli* sequence. The *scale bar* represents the number of inferred nucleotide substitutions per site. *Values at nodes* indicate >50% percentage occurrence in 1,000 bootstrapped trees
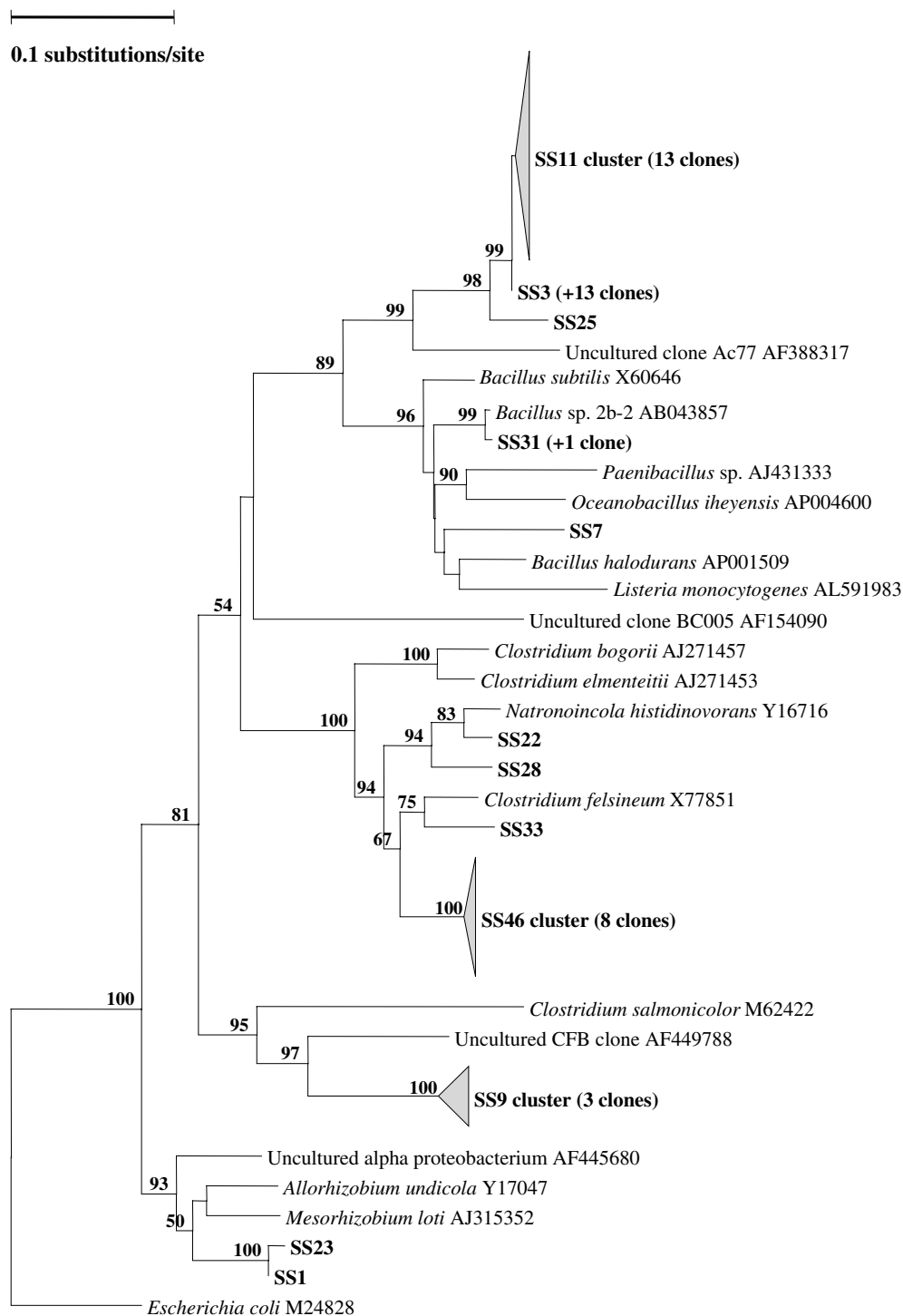
derived clones had been obtained. However, all of the inserts were almost identical in sequence over common regions, indicating that the same cellulase gene was present in each. Three clones contained an ORF of 634 amino acids. The sequence of one of the clones, SScel24, has been deposited in the databases (accession number AJ622825). Nine of the clones contained an ORF of 339 residues, running from amino acids 296–634 of SScel24. BLASTN analysis did not identify any sequences with significant similarity. BLASTP analysis identified an endoglucanase F precursor from *Fibrobacter succinogenes* (accession U39070). This is a large protein of 1,053 amino acid residues. The alignment between the two proteins is over the full length of the SScel24 cellulase and the C-terminal half of the *Fibrobacter* cellulose from residues 403–1,053, as shown in Fig. 2. The identity is 50%, with positive matches being 67%. The *Fibrobacter* cellulase has been experimentally shown to contain an

N-terminal cellulose-binding domain and a C-terminal cellulase domain (Malburg et al. 1997). An NCBI conserved-domain search of the SS cellulase sequence shows significant alignment to three protein families, carbohydrate-binding domain family 11 (pfam 03425), the α-L-arabinofuranosidase domain (COG 3534), and the glycosyl-hydrolase family 79 domain (pfam 03662). These three domains are also found in the *Fibrobacter* protein sequence. Clearly the cellulase activity resides at the C-terminus of SS cellulase from amino acids 296–634, which corresponds to the extreme C-terminus of the *Fibrobacter* protein.

16S rDNA analysis was carried out on clones derived from the same pooled DNA used to make the LS and SS λ libraries and yielded partial sequences of about 450 nucleotides from the 5′ end of the gene. These gave the phylogenetic trees shown in Figs. 3 and 4, respectively. This study, albeit confined to enrichments, is the first molecular analysis of microbial diversity in the Wadi el Natrun lakes.

As expected, a number of the clones from both libraries show closest identity to organisms cultured from saline or alkaline environments, with many representatives of the *Clostridium/Bacillus* group (Jones et al.1998; Duckworth et al.1996).

More than half of the clones from the SS enrichments, the SS3 and SS11 groups, are closely related to each other, and tree with members of the genus *Bacillus*. The closest match in a BLAST search was to an uncultured phylotype from soil in a fertilizer field experiment (accession AF388317). The closest named species was *Paenibacillus* sp. Ikaite C7, isolated from a cold alkaline tufa column in Greenland (Stougaard et al. 2002). Interestingly, the related clone SS25 gets the highest score (E-value 2e-93) in the BLAST search to *Oceanobacillus iheyensis* HTE831, a novel, extremely halotolerant and alkaliphilic species isolated from a depth of 1,050 m on the Iheya Ridge (Lu et al. 2001).

In addition to this major group of closely related clones, several other clades that are apparently distinct from known cultured species can be distinguished. For example, members of the SS9 clade show closest match (E-value 9e-90, 88% identity) to uncultured *Cytophaga–Flexibacter–Bacteroides* (CFB) group bacterial clones from the alkaline hypersaline Mono Lake, California (accession AF449785 and similar clones). Another major clade of nine clones (SS33 and the SS46 group) shows relatedness to *Clostridium felsineum* at around 93% identity.

In the case of the LS enrichments, an even wider diversity of clones was found, perhaps reflecting that several enrichments were pooled in order to obtain sufficient material. Again, the closest matches to some clones were frequently cultivated species (particularly *Bacillus* or *Clostridium*) isolated from alkaline or highly saline environments, while others appeared to be closest to uncultured environmental clones. A larger number than for the SS enrichment library, 13 clones as compared to three, were related to CFB group bacteria.

In all, these enrichment cultures show the presence of a wide diversity of species, often most closely related to as yet uncultured organisms. This demonstrates that DNA from enrichment cultures, as well as direct environmental DNA, can be a valuable source of novel enzymes and biologically active compounds. One important point is that the enzyme activities actually recovered from these enrichment libraries were directed by endocellulase genes, these activities being only a part of the multiple enzymic activities that can drive crystalline cellulose degradation under anaerobic conditions (others include exocellulase and β-glucosidase). Therefore, the other components of the cellulolytic system that actually start the degradation process must also have been present in the libraries, but were not found in this study because of assay bias. One or more of the Gram-positive species whose signatures are numerous in the 16S clone libraries may contribute the other required exocellulase activities.

References

Altschul SF, Hadden TL, Schaffer AA, Zhang J, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402

Amann R, Ludwig W, Schleifer K (1995) Phylogenetic identification and in situ detection of individual microbial cells without cultivation. Microbiol Rev 59:143–169

Duckworth AW, Grant WD, Jones BE, van Steenberg R (1996) Phylogenetic diversity of soda lake alkaliphiles. FEMS Microbiol Ecol 19:181–191

Grant S, Grant WD, Jones BE, Kato C, Li L (1999) Novel archaeal phylotypes from an East African alkaline saltern. Extremophiles 3:139–145

Henne A, Schmitz RA, Bomeke M, Gottschalk G, Daniel R (2000) Screening of environmental DNA libraries for the presence of genes conferring lipolytic activity on *Escherichia coli*. Appl Environ Microbiol 66:3113–3116

Horikoshi K (1999) Alkaliphiles: some applications of their products for biotechnology. Microbiol Mol Biol Rev 63:735–750

Imhoff JF, Sahl HG, Soliman GSH, Truper HG (1979) The Wadi Natrun: chemical composition and microbial mass developments in alkaline brines of eutrophic desert lakes. Geomicrobiol J 1:219–234

Jones BE, Grant WD, Duckworth AW, Owenson GG (1998) Microbial diversity of soda lakes. Extremophiles 2:191–200

Jukes TH, Cantor CR (1969) Evolution of protein molecules In: Munro HN (ed). Mammalian protein metabolism. Academic Press, New York, pp 21–132

Lu J, NogiY, Takami H (2001) *Oceanobacillus iheyensis* gen. nov., sp. nov., a deep-sea extremely halotolerant and alkaliphilic species isolated from a depth of 1,050 m on the Iheya Ridge. FEMS Microbiol Lett 205:291–297

Malburg SR, Malburg LM Jr, Liu T, Iyo AH, Forsberg CW (1997) Catalytic properties of the cellulose-binding endoglucanase F from *Fibrobacter succinogenes* S85. Appl Environ Microbiol 63:2449–2453

Margesin R, Schinner F (2001) Potential of halotolerant and halophilic microorganisms for biotechnology. Extremophiles 5:73–83

Marrs B, Delagrave S, Murphy D (1999) Novel approaches for discovering industrial enzymes. Curr Opin Microbiol 2:241–245

Pfennig N, Lippert KD (1966) Uber das Vitamin B_{12}-Bedurfnis phototropher Schwefelbakterien. Arch Mikrobiol 55:245–256

Rees HC, Grant S, Jones BE, Grant WD, Heaphy S (2003a) Detecting cellulase and esterase activities encoded by novel genes present in environmental DNA libraries. Extremophiles 7:415–421

Rees HC, Jones BE, Grant WD, Heaphy S (2003b) Diversity of Kenyan soda lake alkaliphiles assessed by molecular methods. Extremophiles 8:63–71

Stougaard P, Jorgensen F, Johnsen MG, Hansen OC (2002) Microbial diversity in ikaite tufa columns: an alkaline, cold ecological niche in Greenland. Environ Microbiol 4:487–493

Taher AG (1999) Inland saline lakes of Wadi El Natrun depression, Egypt. Int J Salt Lake Res 8:149–170

Teather RM, Wood PJ (1982) Use of Congo Red-polysaccharide interactions in enumeration and characterisation of cellulolytic bacteria from the bovine rumen. Appl Environ Microbiol 43:777–780

Van de Peer Y, De Wachter R (1994) TREECON: a software package for the construction and drawing of evolutionary trees. Comput Appl Biosci 9:177–182