Sequencing the 500-kb GC-rich Symbiotic Replicon of *Rhizobium* sp. NGR234 Using Dye Terminators and a Thermostable "Sequenase": a beginning

FREIBERG, Christoph, et al.

**Abstract**

Genomes of the soil-borne nitrogen-fixing symbionts of legumes [Azo(Brady)Rhizobium species] typically have GC contents of 59-65 mol%. As a consequence, compressions (up to 400 per cosmid) are common using automated dye primer shotgun sequencing methods. To overcome this difficulty, we have exclusively applied dye terminators in combination with a thermostable "sequenase" for shotgun sequencing GC-rich cosmids from pNGR234a, the 500-kbp symbiotic replicon of Rhizobium sp. NGR234. A thermostable sequenase incorporates dye terminators into DNA more efficiently than Taq DNA polymerase, thus reducing the concentrations needed (20- to 250-fold). Unincorporated dye terminators can simply be removed by ethanol precipitation. Here, we present data of pXB296, one of 23 overlapping cosmids representing pNGR234a. We demonstrate that the greatly reduced number of compressions results in a much faster assembly of cosmid sequence data by comparing assembly of the shotgun data from pXB296 and the data from another pNGR234a cosmid (pXB110) sequenced using dye primer methods. Within the 34,010-bp sequence from pXB296, 28 coding regions [...]
Sequencing the 500-kb GC-rich Symbiotic Replicon of \textit{Rhizobium} sp. NGR234 Using Dye Terminators and a Thermostable "Sequenase": A Beginning

Christoph Freiberg, Xavier Perret, William J. Broughton, and André Rosenthal

Institut für Molekulare Biotechnologie, 07745 Jena, Germany; Laboratoire de Biologie Moléculaire des Plantes Supérieures, Université de Genève, Genève, Switzerland

Genomes of the soil-borne nitrogen-fixing symbionts of legumes [\textit{Azorhizobium} \textit{Bradyrhizobium} \textit{Rhizobium} species] typically have GC contents of 59–65 mol\%. As a consequence, compressions (up to 400 per cosmid) are common using automated dye primer shotgun sequencing methods. To overcome this difficulty, we have exclusively applied dye terminators in combination with a thermostable "sequenase" for shotgun sequencing GC-rich cosmids from pNGR234a, the 500-kbp symbiotic replicon of \textit{Rhizobium} sp. NGR234. A thermostable sequenase incorporates dye terminators into DNA more efficiently than \textit{Taq} DNA polymerase, thus reducing the concentrations needed (20- to 250-fold). Unincorporated dye terminators can simply be removed by ethanol precipitation. Here, we present data of pXB296, one of 23 overlapping cosmids representing pNGR234a. We demonstrate that the greatly reduced number of compressions results in a much faster assembly of cosmid sequence data by comparing assembly of the shotgun data from pXB296 and the data from another pNGR234a cosmid (pXB110) sequenced using dye primer methods. Within the 34,010-bp sequence from pXB296, 28 coding regions were predicted. All of them showed significant homologies to known proteins, including oligopeptide permeases, an essential cluster for nitrogen fixation, and the $C_4$-dicarboxylate transporter DctA.

Soil bacteria of the genera \textit{Azorhizobium}, \textit{Bradyrhizobium}, and \textit{Rhizobium} establish symbiotic associations with leguminous plants leading to the formation of nitrogen-fixing nodules. Plant exudates (particularly flavonoid compounds) modulate the coordinated expression of many bacterial genes leading to the production of mitogenic Nod factors, which provoke nodule formation (Fellay et al. 1995a). In \textit{Azorhizobium} and \textit{Bradyrhizobium} species, symbiotic loci are carried on the chromosome, whereas in \textit{Rhizobium} species most of them are plasmid-borne (Martinez et al. 1990; Fischer 1994; van Rhijn and Vanderleyden 1995). Complete understanding of legume–\textit{Rhizobium} interactions thus requires a catalog of symbiotic genes and exhaustive analysis of their function.

The broad host-range \textit{Rhizobium} sp. NGR234 carries a plasmid of 500 kbp (pNGR234a), which confers on \textit{Agrobacterium tumefaciens} recipients the ability to nodulate certain legumes (Broughton et al. 1984). Identification of symbiotic pNGR234a genes has been performed by host-range extension (complementation) and mutagenesis (Broughton et al. 1986; Lewin et al. 1990), techniques that are not only time-consuming but that fail to identify subtle phenotypes. To facilitate analysis of this replicon, a canonical ordered cosmid library of pNGR234a was constructed (Perret et al. 1991). Flavonoid-inducible loci were mapped to discrete \textit{XhoI} fragments on these cosmids by competitive RNA hybridization (Fellay et al. 1995b). Other symbiotic loci were identified by subtractive DNA hybridization against the genome of the closely related strain \textit{Rhizobium fredii} USDA257 (Perret et al. 1994). Nevertheless, a number of symbiotic genes remain to be identified, and for this reason, we wish to establish the complete nucleotide sequence of pNGR234a.
Automated fluorescent methods have been used to sequence cosmids from eukaryotic organisms, including *Saccharomyces cerevisiae* (Levy 1994), *Caenorhabditis elegans* (Sulston et al. 1992), *Drosophila melanogaster* (Hartl and Palazzolo 1993), and *Homo sapiens* (Bodmer 1994), as well as chromosomes from the prokaryotes *Haemophilus influenzae* (Fleischmann et al. 1995) and *Mycoplasma genitalium* (Fraser et al. 1995). In most large-scale sequencing centers, including ours, this technology is based mainly on the shotgun approach. After random fragmentation of DNA [e.g., cosmids, bacterial artificial chromosomes (BACs), entire chromosomes] using sonication or mechanical forces, size-selected fragments are subcloned into M13 phages, phagemids, or plasmids and sequenced by cycle sequencing using dye primers (Craxton 1993). A disadvantage of this method is that DNA regions with elevated GC contents produce large numbers of compressions (unresolvable foci in sequence gels) leading to several hundred compressions per assembled cosmid sequence. It is known that the use of dye terminators—fluorescently labeled dideoxynucleosidetriphosphates—instead of dye primers reduces the number of compressions (Rosenthal and Charnock-Jones 1993). Therefore, dye terminators are frequently being used for gap closure and proofreading after assembly of the shotgun data.

To sequence GC-rich cosmids with the highest accuracy, we have investigated the effectiveness of shotgun sequencing with dye terminators in comparison to dye primer sequencing. To improve the incorporation of dye terminators into DNA, we have used a modified *Taq* DNA polymerase carrying a single mutation (Tabor and Richardson 1995). This enzyme has properties similar to a thermostable "sequenase" and is commercially available as Thermo Sequenase (Amersham) or AmpliTaq FS (Perkin-Elmer). Concentrations of dye terminators needed in the cycle sequencing reactions can be reduced by 20-250 times. We have found that dye terminator shotgun sequencing leads to compression-free raw data that can be assembled much faster than shotgun data mainly obtained by dye primer sequencing. This strategy thus allows a severalfold increase in speed to sequence individual cosmids. We demonstrate this by comparing assembly of the sequence data of two cosmids from pNGR234a generated by different chemistries: Cosmid pXB296 was sequenced with dye terminators, whereas data for pXB110 were obtained using the common dye primer method. Furthermore, we present the analysis of the entire pXB296 sequence.

**RESULTS**

Comparison of Fluorescent Traces Created by Different Cycle Sequencing Methods

When using a thermostable sequenase [Thermo Sequenase (Amersham)], we were able to reduce the concentrations of dye terminators (Perkin-Elmer) by 20- to 250-fold in comparison to the concentrations needed for *Taq* DNA polymerase without compromising the quality of the sequencing results (Table 1).

To compare the dye terminator and dye primer cycle sequencing procedures, representative templates derived from the pXB296 library were sequenced by both methods, each performed with Thermo Sequenase and *Taq* DNA polymerase (Fig. 1). In general, dye terminator traces do not contain the many compressions (on average, one compression every 50 bases in single reads) that are common with dye primers if mixes do not contain nucleotide analogs like deoxyino-

---

**Table 1. Concentrations (in µM) of Dye Terminators in Each Cycle Sequencing Reaction with Two Different Thermostable DNA Polymerases**

| Dye terminator | AmpliTaq DNA polymerase | Thermo Sequenase DNA polymerase | Dilution factor for dye terminators*
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A <em>Taq</em></td>
<td>0.751</td>
<td>0.017</td>
<td>40</td>
</tr>
<tr>
<td>C <em>Taq</em></td>
<td>22.500</td>
<td>0.137</td>
<td>160</td>
</tr>
<tr>
<td>G <em>Taq</em></td>
<td>0.200</td>
<td>0.009</td>
<td>20</td>
</tr>
<tr>
<td>T <em>Taq</em></td>
<td>45.000</td>
<td>0.183</td>
<td>250</td>
</tr>
</tbody>
</table>

*Thermo Sequenase vs. AmpliTaq.*

---

GENOME RESEARCH - 591
Figure 1  Comparison of sequences from pXB296 created by different cycle sequencing methods. The graphic outputs (program XGAP) of four electropherograms (traces) with the corresponding sequences generated by automatic base calling are shown around the pXB296 sequence position 15,783 (vertical line in the middle of each sequence). The readings cover a part of the minus strand of pXB296. (▼) Extremely low signals produced with dye terminators; (→) the sequence GCA, which is compressed in dye primer scans; (●) automatic base-calling inaccuracies.

sine or 7-deaza-deoxyguanosine triphosphates or if sequencers are used without active heating systems. In addition, dye terminator traces obtained with Thermo Sequenase show more uniform signal intensities over those obtained with Taq DNA polymerase, thus resulting in a reduced number of weak and missing peaks (e.g., a weak G-signal following an A-signal in Thermo Sequenase traces or a weak C-signal following a G-signal in Taq DNA polymerase traces). Using Applied Biosystems (ABI) 373A sequencers, errors in automatic base-calling of Thermo Sequenase/dye terminator scans only arise after 300–350 bases. The average number of resolved bases in dye primer gels (378) is 46 bases longer than in those produced with dye terminators (332 bases). Furthermore, in Thermo Sequenase/dye primer sequences the peaks are very regular and the number of stops and missing bases decreases in comparison to Taq DNA polymerase/dye terminator electropherograms. The number of compressions, however, is not significantly reduced.

Shotgun Sequencing of Entire Cosmids Using Dye Terminators or Dye Primers

To compare the efficiency of both methods, cosmid pXB296 of pNGR234a was shotgun sequenced using a combination of dye terminators and thermostable sequenase (Thermo Sequenase), whereas another cosmid, pXB110, was sequenced using a combination of dye primers and Taq DNA polymerase (Table 2). Over 93% (736 clones) of 786 dye terminator reads of pXB296 were accepted by XGAP with a maximal alignment mismatch of 4%. By increasing this level to 25%, so that most of the remaining data could be included in the assembly, 775 reads led to three 6- to 10-kbp stretches of contiguous sequence (contigs), two of which were joined after editing. To close the last gap and to complete single-stranded regions with data derived from the opposite strand, only 32 additional dye terminator reads using custom-made primers were obtained. It took <1 week to assemble and finalize the 34,010-bp DNA sequence of pXB296 (EMBL accession no. Z68203; eightfold redundancy; GC content, 58.5 mol%).

In contrast, only 308 (34%) of 899 shotgun reads obtained by dye primer/Taq DNA polymerase cycle sequencing of pXB110 were included in the first assembly (4% alignment mismatch). At the 25% alignment mismatch level, 879 reads were assembled, leading to 25 short contigs (1–2
kbp). These ("contigs") had to be edited extensively in order to join most of them. "Primer walks" covering gaps and complementing single-stranded regions, were not sufficient to clarify all of the remaining ambiguities in the assembled sequence. Every 100–150 bp, a compression in one strand could not be resolved by sequence data from the complementary strand. Therefore, it was necessary to resequence clones using dye terminators and universal primer. In total, 191 additional dye terminator reads had to be created. As a result, assembling and finalizing the 34,573-bp sequence of pXB110 (10.5-fold redundancy; GC content, 58.3 mol%) took much more time than pXB296 did.

Analysis of Cosmid pXB296

Putative genes were located on the 34,010-bp sequence of pXB296 using the programs TESTCODE (Fickett 1982) and CODONPREFERENCE (Gribskov et al. 1984), the latter in combination with a codon frequency table based on previously sequenced genes of Rhizobium sp. NGR234 (as well as the closely related R. fredii). All 28 open reading frames (ORFs) and their deduced amino acid sequences exhibited significant homologies to known genes and/or proteins. The positions of the ORFs along pXB296, as well as the best homologs, are displayed in Table 3 and Figure 2. Ribosomal binding site-like sequences (Shine and Dalgarno 1974) precede each putative gene except for ORF9 (position 11,124–12,455). If one disregards the homology to known glutamate dehydrogenases in the first 32 amino acids deduced from this ORF, a downstream alternative start codon (position 11,220) preceded by a Shine–Dalgarno sequence can be identified. Most of the ORFs are organized in five clusters (ORFs with only short intergenic spaces or overlaps between them). Cluster I containing ORF1 to ORF5 encodes proteins homologous to trans-membrane and membrane-associated oligopeptide permease proteins and to a Bacillus anthracis encapsulation protein. Cluster II includes ORF6 and ORF7 homologous to aminotransferase and (semi)aldehyde dehydrogenase genes. Homologies to transposase genes [ORF8; cluster III (ORF10 and ORF11)] and to various nif and fix genes [cluster IV (ORF12 to ORF20); ORF23, part of cluster V] are also reported.

Presumed promoter and stem–loop sequences that might represent ρ-independent terminator-like structures (Platt 1986) are shown in Figure 2. Significant α84-dependent promoter consensus sequences (5'-GATCAG-N4-TTGCA-3'; Morett and Buck 1989), as well as nifA upstream activator sequences (5'-TGT-N10-ACA-3';

<table>
<thead>
<tr>
<th>Table 2. Comparison of the Assembly of the Sequence Data from Cosmids pXB296 (Dye Terminator Shotgun Reads) and pXB110 (Dye Primer Shotgun Reads)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data assembly</td>
</tr>
<tr>
<td>----------------</td>
</tr>
<tr>
<td>Average length of the shotgun reads (bases)</td>
</tr>
<tr>
<td>No. of shotgun reads used for assembly</td>
</tr>
<tr>
<td>No. of shotgun reads assembled with 4% mismatcha</td>
</tr>
<tr>
<td>No. of shotgun reads assembled with 25% mismatcha</td>
</tr>
<tr>
<td>No. of contigsb longer than 1 kbp</td>
</tr>
<tr>
<td>No. of contigs left after editingc</td>
</tr>
<tr>
<td>No. of additional reads (gap closure and proofreading)d</td>
</tr>
<tr>
<td>Total length of cosmid insert (bp)</td>
</tr>
<tr>
<td>Sequencing redundancy (per bp)</td>
</tr>
</tbody>
</table>

aAssembling program: XGAP; principal autoassembling conditions: normal shotgun assembly, joins permitted, minimum initial match = 15, maximum no. of pads per reading during the alignment procedure = 8, maximum no. of pads per reading in contig to align any new reading = 8, alignment mismatches 4% and 25%, respectively.

bLengths of contigs: 6–10 kbp (pXB296); 2–12 kbp (pXB110).

cContiguous parts of sequence created by overlapping reads.

dReads necessary for closing gaps and making single-stranded regions double-stranded by primer walking on selected templates and, in case of pXB110, for solving ambiguities (compressions) by the resequencing of clones with universal primer and dye terminators.

GENOME RESEARCH 179
Morett and Buck 1988), are found upstream of the \( \text{nifB} \) homolog ORF15, the \( \text{fixA} \) homolog ORF20, and ORF21, ORF22, and ORF23. ORF23 is part of cluster V in pXB296, which includes the \( \text{dctA} \) gene of Rhizobium sp. NGR234 (van Slooten et al. 1992). Surprisingly, the published \( \text{dctA} \) sequence shows important discrepancies. Therefore, a fragment encompassing this locus was amplified by PCR using NGR234 genomic DNA as template. By sequencing this fragment, our cosmid sequence was confirmed.

**DISCUSSION**

Advantages of the Dye Terminator/Thermostable Sequenase Shotgun Strategy

We have examined whether GC-rich cosmids can be sequenced much more efficiently using dye

### Table 3. Putative Genes of pXB296 and Homologies of the Deduced Amino Acid Sequences to Known Proteins

<table>
<thead>
<tr>
<th>ORF</th>
<th>Start Position (base no.)</th>
<th>Ribosomal Binding Site</th>
<th>SD-Sequence Distance from Start Codon</th>
<th>No. of Deduced Amino Acids</th>
<th>Homologous Protein</th>
<th>Identity (%)</th>
<th>Similarity (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ORF1</td>
<td>+ 00001-00625</td>
<td>5’à-TAGGAGGTTG-3’à</td>
<td>&gt;207</td>
<td>1-207</td>
<td>OppB</td>
<td>306</td>
<td>oligopeptide</td>
</tr>
<tr>
<td>ORF2</td>
<td>+ 00628-01503</td>
<td>GTATCCGGTCG-7´-ATG</td>
<td>291</td>
<td>2-289</td>
<td>OppC</td>
<td>305</td>
<td>permease</td>
</tr>
<tr>
<td>ORF3</td>
<td>+ 01505-02512</td>
<td>AGGGAGGACG-7´-ATG</td>
<td>335</td>
<td>8-327</td>
<td>OppD</td>
<td>336</td>
<td>permease</td>
</tr>
<tr>
<td>ORF4</td>
<td>+ 02509-03570</td>
<td>TCAGATGGGT-6´-ATG</td>
<td>353</td>
<td>2-323</td>
<td>OppF</td>
<td>334</td>
<td></td>
</tr>
<tr>
<td>ORF5</td>
<td>+ 03606-04991</td>
<td>CAAGGA-6´-ATG</td>
<td>461</td>
<td>1-458</td>
<td>CupA</td>
<td>411</td>
<td>encapsulation protein</td>
</tr>
<tr>
<td>ORF6</td>
<td>+ 05460-06863</td>
<td>CCGAGAGG-8´-ATG</td>
<td>467</td>
<td>1-464</td>
<td>BioA</td>
<td>455</td>
<td>aminotransferase</td>
</tr>
<tr>
<td>ORF7</td>
<td>+ 06888-08426</td>
<td>GCCCTTCG-5´-GTA</td>
<td>512</td>
<td>97-509</td>
<td>ORF5</td>
<td>417</td>
<td>unknown</td>
</tr>
<tr>
<td>ORF8</td>
<td>- 09781-10860</td>
<td>GACCTCTTCG-8´-ATG</td>
<td>359</td>
<td>72-299</td>
<td>ORF2</td>
<td>414</td>
<td>transposase, homologous, minicircle DNA</td>
</tr>
<tr>
<td>ORF9</td>
<td>+ 11124-12455</td>
<td>?-7´-ATG</td>
<td>443</td>
<td>2-443</td>
<td>GLUD1</td>
<td>558</td>
<td>glutamate dehydrogenase</td>
</tr>
<tr>
<td>ORF10</td>
<td>- 13370-14116</td>
<td>AAAAGG-6´-ATG</td>
<td>248</td>
<td>1-245</td>
<td>ORF2</td>
<td>231</td>
<td>transposase</td>
</tr>
<tr>
<td>ORF11</td>
<td>- 14128-15672</td>
<td>CATGGAGG-7´-ATG</td>
<td>514</td>
<td>1-513</td>
<td>ORF1</td>
<td>558</td>
<td>homologues, IS1162</td>
</tr>
<tr>
<td>ORF12</td>
<td>- 16712-16942</td>
<td>GAGGAGG-8´-ATG</td>
<td>76</td>
<td>1-70</td>
<td>FixU</td>
<td>70</td>
<td>unknown</td>
</tr>
<tr>
<td>ORF13</td>
<td>- 16939-17265</td>
<td>ACAAGAGG-7´-ATG</td>
<td>109</td>
<td>1-79</td>
<td>ORF2</td>
<td>78</td>
<td>unknown, involved in FeMo-cofactor synthesis</td>
</tr>
<tr>
<td>ORF14</td>
<td>- 17349-17543</td>
<td>CCGAGAGG-9´-ATG</td>
<td>64</td>
<td>1-64</td>
<td>FdxN</td>
<td>64</td>
<td>ferredoxin-like</td>
</tr>
<tr>
<td>ORF15</td>
<td>- 17585-19066</td>
<td>AGTGGAGG-7´-ATG</td>
<td>493</td>
<td>1-493</td>
<td>NifB</td>
<td>490</td>
<td>involved in FeMo-cofactor synthesis</td>
</tr>
<tr>
<td>ORF16</td>
<td>- 19292-20562</td>
<td>ATGAGG-12´-ATG</td>
<td>556</td>
<td>9-556</td>
<td>NifA</td>
<td>541</td>
<td>transcriptional regulator</td>
</tr>
<tr>
<td>ORF17</td>
<td>- 21129-21422</td>
<td>ATGGAGG-7´-ATG</td>
<td>97</td>
<td>9-97</td>
<td>FixX</td>
<td>98</td>
<td>required for nitrogen fixation</td>
</tr>
<tr>
<td>ORF18</td>
<td>- 21437-22744</td>
<td>ACTGGAGG-7´-ATG</td>
<td>435</td>
<td>1-435</td>
<td>FixC</td>
<td>435</td>
<td></td>
</tr>
<tr>
<td>ORF19</td>
<td>- 22755-23864</td>
<td>ATGAGG-6´-ATG</td>
<td>369</td>
<td>18-369</td>
<td>FixD</td>
<td>353</td>
<td></td>
</tr>
<tr>
<td>ORF20</td>
<td>- 23874-24731</td>
<td>ATGAGG-5´-ATG</td>
<td>285</td>
<td>1-285</td>
<td>FixA</td>
<td>292</td>
<td></td>
</tr>
<tr>
<td>ORF21</td>
<td>- 25148-25468</td>
<td>CCGAGAGG-10´-ATG</td>
<td>106</td>
<td>1-106</td>
<td>ORF18</td>
<td>108</td>
<td>unknown</td>
</tr>
<tr>
<td>ORF22</td>
<td>- 26145-26711</td>
<td>GAGGAGG-9´-ATG</td>
<td>188</td>
<td>9-199</td>
<td>ORF18</td>
<td>241</td>
<td>hypothetical protein</td>
</tr>
</tbody>
</table>

594 | GENOME RESEARCH
terminators throughout the shotgun phase instead of dye primers. As a test case, cosmid pXB296 with a GC content of 58 mol% from pNGR234a, the symbiotic plasmid of Rhizobium sp. NGR234, was exclusively sequenced using dye terminators in combination with a thermostable sequenase [Thermo Sequenase (Amersham)]. Another rhizobial cosmid with identical GC content, pXBl10, was sequenced using traditional dye primer chemistry and Taq DNA polymerase.

Using the dye terminator/thermostable sequenase shotgun strategy, we have shown that most, if not all, compressions could be resolved and reads were produced with the highest fidelity among all sequencing chemistries tested. As a result, we obtained a much faster assembly of cosmid pXB296 in comparison to pXBl10. The shotgun data could be assembled to a high-quality sequence without extensive editing and proofreading. By measuring the error rate in overlapping regions between individual cosmids from pNGR234a, as well as in the cosmid vector sequence itself (data not shown), we estimate that the accuracy of the pXB296 sequence is higher than 99.98%. Using other thermostable sequenases such as AmpliTaq FS (Perkin-Elmer), we would expect similar results because thermostable sequenases have similar properties.

We also examined dye primer chemistry in combination with Thermo Sequenase. Although the peak uniformity of signals was much improved over dye primer/Taq DNA polymerase data, the number of compressions in GC-rich shotgun reads was not reduced significantly. Compressions in shotgun raw data enormously increase the overall effort of editing, proofreading, and finishing a cosmid as shown for pXBl10 (Table 2). We have not investigated whether compressions in dye primer sequencing of rhizobial cosmids could be reduced efficiently using either nucleotide analogs like deoxyinosine or 7-deaza-deoxyguanosine triphosphates or el-

### Table 3. (Continued)

<table>
<thead>
<tr>
<th>ORF</th>
<th>Start</th>
<th>Stop</th>
<th>Length</th>
<th>NifQ</th>
<th>ATG</th>
<th>Probably Involved Mo-processing</th>
<th>X13303</th>
<th>37</th>
<th>57</th>
</tr>
</thead>
<tbody>
<tr>
<td>ORF23</td>
<td>27169</td>
<td>27861</td>
<td>702</td>
<td>167</td>
<td>ATG</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ORF24</td>
<td>27920</td>
<td>29434</td>
<td>1525</td>
<td>456</td>
<td>ATG</td>
<td></td>
<td>538912</td>
<td>97</td>
<td>98</td>
</tr>
<tr>
<td>ORF25</td>
<td>29431</td>
<td>30675</td>
<td>1245</td>
<td>415</td>
<td>ATG</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ORF26</td>
<td>30676</td>
<td>31332</td>
<td>656</td>
<td>155</td>
<td>ATG</td>
<td></td>
<td>538912</td>
<td>97</td>
<td>98</td>
</tr>
<tr>
<td>ORF27</td>
<td>31329</td>
<td>33035</td>
<td>706</td>
<td>244</td>
<td>ATG</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ORF28</td>
<td>33733</td>
<td>34010</td>
<td>277</td>
<td>355</td>
<td>ATG</td>
<td></td>
<td>538912</td>
<td>97</td>
<td>98</td>
</tr>
</tbody>
</table>

*ORF* Open reading frame.

Position on cosmid: from the first base of the start codon to the last base of the stop codon; alternative start points are 6912/6927/7017 (ORF7), 10665/10656 (ORF8), 11220 (ORF9), 15699/15651 (ORF11), 17322/17271 (ORF13), 20995/21076 (ORF16), 26744 (ORF22), 27229/27304 (ORF23), 27941 (ORF24), and 30751/30754 (ORF26).

*SD sequence* Shine–Dalgarno sequence (Shine and Dalgarno 1974). Bases underlined are identical with the Shine–Dalgarno sequence. The following possible start codons were considered: ATG, GTG, or TGT.

*aa* Amino acids.

*Organisms:* Salmonella typhimurium, Bacillus subtilis (CapA), Bacillus subtilis (BioA), Streptomyces hygroscopicus (ORF7 homolog), Escherichia coli (GapD), Streptomyces coelicolor (ORF8 homolog), Homo sapiens (GLUD1), Pseudomonas fluorescens (ORF10, ORF11 homologs), Rhizobium leguminosarum (FixXU), Rhodobacter capsulatus (ORF13 homolog), Azotobacter vinelandii (Nif2), Rhizobium meliloti (FdxN, NifB, FixXCA), Bradyrhizobium japonicum (ORF118), Haemophilus influenzae (hypothetical protein), Lipomyces kononenkoi (peroxisomal protein), Klebsiella pneumoniae (NifQ), Rhizobium sp. NGR234 (DctA), Pseudomonas putida (CapC), Pseudomonas putida (LineA), Escherichia coli (FabG), Vibrio harveyi (LuxA).

Identity and similarity were calculated using the program BESTFIT (Smith and Waterman 1981).

Table 3. (Continued)

<table>
<thead>
<tr>
<th>ORF</th>
<th>Start</th>
<th>Stop</th>
<th>Length</th>
<th>NifQ</th>
<th>ATG</th>
<th>Probably Involved Mo-processing</th>
<th>X13303</th>
<th>37</th>
<th>57</th>
</tr>
</thead>
<tbody>
<tr>
<td>ORF23</td>
<td>27169</td>
<td>27861</td>
<td>702</td>
<td>167</td>
<td>ATG</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ORF24</td>
<td>27920</td>
<td>29434</td>
<td>1525</td>
<td>456</td>
<td>ATG</td>
<td></td>
<td>538912</td>
<td>97</td>
<td>98</td>
</tr>
<tr>
<td>ORF25</td>
<td>29431</td>
<td>30675</td>
<td>1245</td>
<td>415</td>
<td>ATG</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ORF26</td>
<td>30676</td>
<td>31332</td>
<td>656</td>
<td>155</td>
<td>ATG</td>
<td></td>
<td>538912</td>
<td>97</td>
<td>98</td>
</tr>
<tr>
<td>ORF27</td>
<td>31329</td>
<td>33035</td>
<td>706</td>
<td>244</td>
<td>ATG</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ORF28</td>
<td>33733</td>
<td>34010</td>
<td>277</td>
<td>355</td>
<td>ATG</td>
<td></td>
<td>538912</td>
<td>97</td>
<td>98</td>
</tr>
</tbody>
</table>

*ORF* Open reading frame.

*Position on cosmid: from the first base of the start codon to the last base of the stop codon; alternative start points are 6912/6927/7017 (ORF7), 10665/10656 (ORF8), 11220 (ORF9), 15699/15651 (ORF11), 17322/17271 (ORF13), 20995/21076 (ORF16), 26744 (ORF22), 27229/27304 (ORF23), 27941 (ORF24), and 30751/30754 (ORF26).

*SD sequence* Shine–Dalgarno sequence (Shine and Dalgarno 1974). Bases underlined are identical with the Shine–Dalgarno sequence. The following possible start codons were considered: ATG, GTG, or TGT.

*aa* Amino acids.

*Organisms:* Salmonella typhimurium, Bacillus subtilis (OppBCDF), Bacillus subtilis (CapA), Bacillus subtilis (BioA), Streptomyces hygroscopicus (ORF7 homolog), Escherichia coli (GapD), Streptomyces coelicolor (ORF8 homolog), Homo sapiens (GLUD1), Pseudomonas fluorescens (ORF10, ORF11 homologs), Rhizobium leguminosarum (FixU), Rhodobacter capsulatus (ORF13 homolog), Azotobacter vinelandii (Nif2), Rhizobium meliloti (FdxN, NifB, FixXCA), Bradyrhizobium japonicum (ORF118), Haemophilus influenzae (hypothetical protein), Lipomyces kononenkoi (peroxisomal protein), Klebsiella pneumoniae (NifQ), Rhizobium sp. NGR234 (DctA), Pseudomonas putida (CapC), Pseudomonas putida (LineA), Escherichia coli (FabG), Vibrio harveyi (LuxA).

Identity and similarity were calculated using the program BESTFIT (Smith and Waterman 1981).

*ORF* 3' end.

*Translated ORF.

*ORF* 5' end.

Sequencing the Symbiotic Replicon of Rhizobium
evated gel running temperatures. Because of their longer
reading potential, dye primer reads are definitely helpful for
gap closure. However, using ABI 373A sequencers, dye
primer reads are, on average, only ~50 bases longer than dye
terminator reads.

We are aware that from the methodological point of
view our comparison is not quite parallel because two dif­
erent cosmids were compared using two different enzymes.
However, under our experimental conditions, shotgun se­
quencing with dye terminators and a thermostable sequen­
sing enzyme is superior because for GC-rich

Figure 2 Organization of the predicted genes in pXB296 from Rhizobium
sp. NGR234. Significant stem–loops (/) and stem–loop clusters (■), which might
function as ρ-independent transcription terminators, are represented. Se­
quence motifs (open arrowhead), similar to σ54-dependent promoter con­
sensus sequences (GGCACG–N₄–TIGC + nifA upstream activator sequences
(TGT–N₁₀–ACA), are located at the following positions on the cosmid:
19,107–19,120 + 19,195–19,210; 24,787–24,800 + 24,878–24,893;
25,508–25,521 + 25,567–25,582; 26,815–26,828 + 26,941–26,956 (all minus strand); 27,074–27,131 (several possibilities) + 26,969–26,984 (plus strand).

Genetic Organization of pXB296
All 28 predicted ORFs in pXB296 (Fig. 2) show
significant homologies to database entries. The
first putative gene cluster (cluster I) containing
ORF1 to ORF5 corresponds to various oligopep­
tide permease operons (Hiles et al. 1987; Perego
et al. 1991). Only ORF5 shows homology to a
gene from a different bacterium, B. anthracis
(Makino et al. 1989). Each homolog encodes
membrane-bound or membrane-associated pro­
teins suggesting that all five genes are involved in
oligopeptide permeation.

Organization of the predicted gene cluster
IV, including the nifA homolog ORF16 (fixABCX,
nifA, nifB, fixN, ORF, fixU) homologs, position
16,746–24,731), as well as the predicted locations of the σ54-dependent promoters and the nifA up­
stream activator sequences (Fig. 2), corresponds to the organization found in Rhizobium meliloti and
Rhizobium leguminosarum bv. trifolii. (Iismaa et al. 1989; Fischer 1994). NifA is a positive tran­
scriptional activator (Buikema et al. 1985), whereas nif and fix genes are essential for sym­ionic nitrogen fixation. Identification of σ54­
dependent promoter sequences, together with
the upstream activator motifs upstream of
ORF21, ORF22, and ORF23 suggests that these
genes may play an important, but still undefined
role, in symbiosis.

Inevitably, large-scale sequencing uncovers
differences with already published sequences.
van Slooten et al. (1992) cloned a 5.8-kb EcoRI
fragment from Rhizobium sp. NGR234 and se­
quenced 2067 bp by manual radioactive methods
(EMBL accession no. S38912). This sequence ex­
hits 2.4% mismatches with the corresponding
sequence in pXB296. It contains the gene dctA
(encoding a C₄-dicarboxylate permease), which is
144 bases shorter than in pXB296. In this respect,
a single nucleotide deletion in position 29,248 of
the cosmid sequence close to the 3’ end of the
gene causes a frameshift leading to a DctA prod­
cut extended by 48 residues. van Slooten et al.
(1992) also failed to identify the nifQ homolog,
ORF23 (position 27,169–27,861), presumably be-
cause they overlooked a small XhoI fragment located between positions 27,349 and 27,536 on pXB296. Expression studies allowed these investigators to define a putative promoter in a 1.7-kb SmaI fragment (position 27,094–28,818 in pXB296). This fragment stretches from the upstream region of ORF23 to the 5′ part of dctA. The 58-bp intergenic region between ORF23 and dctA contains a stem–loop structure but no obvious promoter sequence. Possibly the promoter that controls dctA is located upstream of ORF23 (e.g., the minimal consensus sequence included in GGGGGCCAATTTGCG at position 27,098–27,111). Although clones containing dctA complemented mutants of R. meliloti and R. leguminosarum for growth on dicarboxylates, the growth of the NGR234 dctA deletion mutant was not affected (van Slooten et al. 1992). Nevertheless, this mutant was unable to fix nitrogen in nodules. Because dctA is now possibly part of a larger transcription unit, the symbiotic phenotype may also result from the inactivation of downstream genes.

Interestingly, the GC content of the predicted pXB296 ORFs ranges from 53.3 mol% to 64.6 mol%, with an overall cosmid GC content of 58.5 mol%. Genomes of Azorhizobium, Bradyrhizobium, and Rhizobium species have GC contents of 59 mol% to 65 mol% (Padmanabhan et al. 1990), with 62 mol% reported for Rhizobium sp. NGR234 (Broughton et al. 1972). Although pXB296 covers <7% of the complete symbiotic genome, its lower overall GC value suggests that symbiotic genes might have evolved by lateral transfer from other organisms. In this case, methods of the type applied here will become even more relevant in sequencing the whole genome.

Although functional analyses of selected ORFs in pXB296 still have to be performed, large-scale sequencing gives a global picture of their genomic organization and possible roles. Determination of putative functions of predicted genes by homology searches and identification of sequence motifs (promoters, nod boxes, nifA activator sequences, and other regulatory elements) will aid in finding new symbiotic genes. High-fidelity sequence data covering long stretches of the genome are a prerequisite for these studies. The dye terminator/thermostable sequenase shotgun approach will allow completion of the entire 500-kb sequence of pNGR234a within several months and open up new avenues for genetic analysis of symbiotic function.

**METHODS**

**Bacteria and Plasmids**

_Escherichia coli_ was grown on SOC, in TB, or in twofold YT medium (Sambrook et al. 1989). The cosmid clones pXB296 and pXB110 (Perret et al. 1991) were raised in _E. coli_ strain 1046 (Cami and Kourilsky 1978). Subclones in M13mp18 vectors (Tanisch-Perron et al. 1985) were grown in _E. coli_ strain DH5αFIQ (Hanahan 1983).

**Construction of pXB296 and pXB110 Libraries**

Cosmid DNA was prepared by standard alkaline lysis procedures followed by purification in CsCl gradients (Radloff et al. 1967). DNA fragments sheared by sonication of 10 µg of cosmid DNA were treated for 10 min at 30°C with 30 units of mung bean nuclease (New England Biolabs, Beverly, MA), extracted with phenol/chloroform (1:1), and precipitated with ethanol. DNA fragments, ranging in size from 1 to 1.4 kbp, were purified from agarose gels using GeneClean II (Biorad, Vista, CA) and ligated into M13mp18. Electroporation of aliquots of the ligation reaction into competent _E. coli_ DH5αFIQ was performed according to standard protocols (Dower et al. 1988; Sambrook et al. 1989).

**M13 Template Preparation**

Fresh 1-ml _E. coli_ cultures in twofold YT held in 96-deep well microtiter plates (Beckman Instruments, Fullerton, CA) were infected with recombinant phages from white plaques growing on plates containing X-gal (5-bromo-4-chloro-indoly-β-D-galactoside) and IPTG (isopropyl-1-β-D-thio-galactopyranoside). Rapid preparation of ~0.5 µg of single-stranded M13 template DNA was carried out as follows: 190-µl portions of the phage cultures grown for 6 hr at 37°C were transferred into 96-well microtiter plates. Lysis of the phages was obtained by adding 10 µl of 15% (wt/vol) SDS followed by 5 min incubation at 80°C. Template DNA was trapped using 10 µl (1 mg) of paramagnetic beads (Streptavidin MagnSphere Paramagnetic Particles Plus M13 Oligo, Promega, Madison, WI) and 50 µl of hybridization solution [2.5 mM NaCl, 20% (wt/vol) polyethylene glycol (PEG-8000)] during an annealing step of 20 min at 45°C. Beads were pelleted by placing microtiter plates on appropriate magnets and washed three times with 100 µl of 0.1-fold SSC. The DNA was recovered in 20 µl of water by a denaturation step of 3 min at 80°C. When required, larger amounts of single-stranded recombinant DNA (>10 µg) were purified using QIAprep 8 M13 Purification Kits (Qiagen, Hilden, Germany) from 3 ml of supernatant of phage cultures grown for 6 hr at 37°C.

**Sequencing**

Two sequencing methods were used: dye terminator and dye primer cycle sequencing, each in combination with AmpliTaq DNA polymerase (Perkin-Elmer, Foster City, CA) and Thermo Sequenase (Amersham, Buckinghamshire, UK). All reactions, including ethanol precipitation, were performed in microtiter plates. Reagents were pipetted us-
ing 12-channel pipettes. Where necessary, sequencing reaction mixtures, including enzymes, were pipetted into the plates in advance and held at -20°C until needed.

Dye Terminator Cycle Sequencing

For dye terminator/AmpLiTaq DNA polymerase sequencing, 0.5 µg of template DNA, and the PRISM Ready Reaction DyeDeoxy Terminator Cycle Sequencing Kit (Perkin-Elmer) were used. Cycle sequencing was performed in microtiter plates using 25 PCR cycles (30 sec at 95°C, 30 sec at 50°C, and 4 min at 60°C). Prior to loading the amplified products on electrophoresis gels, unreacted dye terminators were removed using Sephadex columns scaled down to microtiter plates (Rosenthal and Charnock-Jones 1993).

Dye terminator/Thermo Sequenase sequencing was performed using the same experimental conditions except that the reaction mix contained 16.25 mM Tris-HCl (pH 9.5), 4.0 mM MgCl₂, 0.02% (vol/vol) Tween 20, 0.5 µg of template DNA, and the PRISM Ready Reaction mixtures, including enzymes, were pipetted into microtiter plates using 25 PCR cycles (30 sec at 95°C, 30 sec at 50°C, and 4 min at 60°C). Prior to loading the amplified products on electrophoresis gels, unreacted dye terminators were removed using Sephadex columns scaled down to microtiter plates (Rosenthal and Charnock-Jones 1993).

Dye terminator/Thermo Sequenase sequencing was performed using the same experimental conditions except that the reaction mix contained 16.25 mM Tris-HCl (pH 9.5), 4.0 mM MgCl₂, 0.02% (vol/vol) Tween 20, 42 µM 2-mercaptoethanol, 100 µM dATP/dTTP, 300 µM dITP, 0.017 µM A/0.137 µM C/0.009 µM G/0.183 µM T from Taq Dye Terminators (Perkin-Elmer; no. ASF034), 0.67 µM primer, 0.2-0.5 µg of template DNA, and 10 units of Thermo Sequenase (Amersham) in a 30-µl reaction volume. Unincorporated dye terminators were removed from reaction mixtures by precipitation with ethanol.

Dye Primer Cycle Sequencing

Dye primer/AmpLiTaq DNA polymerase sequencing reactions were performed according to the instructions accompanying the Taq Dye Primer, 21M13 Kit (Perkin-Elmer). Cycle sequencing was carried out on 0.5 µg of template DNA with 19 PCR cycles (30 sec at 95°C, 30 sec at 50°C, and 90 sec at 72°C) followed by six cycles, each consisting of 95°C for 30 sec and 72°C for 2.5 min. Prior to electrophoresis, the four base-specific reactions were pooled and precipitated with ethanol.

Identical PCR conditions and the Thermo Sequenase Fluorescent Labeled Primer Cycle Sequencing Kit (Amersham) were used for dye primer/Thermo Sequenase sequencing reactions.

Sequence Acquisition and Analysis

Gel electrophoresis and automatic data collection were performed with ABI 373A DNA sequencers (Perkin-Elmer). After removing cosmid vector and M13mp18 sequences from the shotgun sequence data, the data were assembled using the program XGAP (Dear and Staden 1991) and edited against the fluorescent traces. To close remaining gaps, to make single-stranded regions double-stranded, and to clarify ambiguities, additional cycle sequencing reactions with selected shotgun templates were carried out using either custom-made primers (primer-walks) or universal primer.

The complete double-stranded DNA sequence of cosmid pXB296 was analyzed using programs from the Wisconsin Sequence Analysis Package (v. 8, Genetics Computer Group, Madison, WI). Homology searches were performed with BLAST (v. 1.4; Altschul et al. 1990) and FASTA (v. 2.0; Pearson and Lipman 1988). Several nucleotide and protein data bases were screened (GenBank/Genpept, SwissProt, EMBL, and PIR). Identities and similarities between homologous amino acid sequences were calculated with the alignment program BESTFIT (Smith and Waterman 1981).

ACKNOWLEDGMENTS

We thank Matthias Platzer for help and critical reading of the manuscript, Evelyn MichaeIl for excellent technical assistance, and Bernd Drescher for managing the computer system. X.P. gratefully acknowledges the receipt of a European Molecular Biology Organization (EMBO) short-term fellowship. Financial support was provided by the Fonds National Suisse de la Recherche Scientifique (grant 31-36454.92). The sequence data described in this paper have been submitted to the EMBL data library under accession no. Z68203.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES


Dear, S. and R. Staden. 1991. A sequence assembly and
SEQUENCING THE SYMBIOTIC REPLICON OF RHIZOBIUM


GENOME RESEARCH 599


Received February 22, 1996; accepted in revised form May 23, 1996.