Analysis of pFQ12, a 22.4-kb *Frankia* plasmid

Theodore R. John, Jeffrey M. Rice, and Jerry D. Johnson

Abstract: *Frankia* are gram-positive, filamentous bacteria capable of fixing atmospheric dinitrogen in symbiosis with a wide variety of woody plants and shrubs. Some isolates of *Frankia* harbor plasmids of 8.5 (pFQ11) and 22.4 kb (pFQ12) that have no known function but are transmitted through many generations in culture. We have sequenced the 22 437-bp pFQ12 plasmid that is present in isolates CpI1 and ArI3. This sequence, with 76% G+C, is almost totally unrelated to that of pFQ11 found in the same cells. However, four regions of identity, 40–90 bp each, are dispersed around the plasmids. The 22.4-kb plasmid has >50 open reading frames (ORFs) that encode putative proteins of more than 100 amino acids, with the largest being 2226 amino acids. Twenty of these ORFs are likely to encode proteins based on their codon bias as determined by two different algorithms. Transcripts from nine of these regions have been identified by reverse transcriptase–polymerase chain reaction (RT–PCR) or filter hybridization. The two *Frankia* plasmids each encode a protein similar to the korSA protein that regulates transmission of pSAM2 in *Streptomyces*. The origin of replication (ORI) region of pFQ12 was localized by intrastrand AT and GC equivalence switch. It includes a 40-bp, intergenic, A+T-rich region that has a strong identity in pFQ11.

Key words: ORI analysis, RT–PCR, Glimmer, DNA sequence.

Résumé : *Frankia* est une bactérie filamenteuse gram-positive capable de fixer le diazote atmosphérique lorsqu'en symbiose avec une grande variété de plantes ligneuses et d'arbustes. Certains isolats de *Frankia* contiennent des plasmides de 8,5 (pFQ11) et 22,4 kb (pFQ12) qui n'ont aucune fonction connue mais qui sont transmises à travers plusieurs générations en culture. Nous avons séquencé le plasmide pFQ12 de 22 437 pb qui est présent dans les isolats CpI1 et ArI3. Cette séquence, constituée à 76 % de G+C, est presque totalement non apparentée à celle de pFQ11 retrouvé dans les mêmes cellules. Cependant, quatre régions d'identité, de 40–90 pb chaquene, sont dispersées autour des plasmides. Le plasmide de 22,4 kb contient >50 cadres de lecture ouverts « CLO » dant des protéines putatives de plus de 100 acides aminés, la plus grosse étant composée de 2226 acides aminés. Vingt d'entre ceux-ci sont susceptibles de coder des protéines, en se basant sur leur préférence en codons tel que déterminé par deux algorithmes différents. Les transcrits de neuf de ces régions ont été identifiés par RT–PCR ou par hybridation sur filtre. Les deux plasmides de *Frankia* codent chacune une protéine semblable à la protéine korSA qui contrôle la transmission de pMSA2 dans *Streptomyces*. La région ORI (« origin of replication ») de pFQ12 a été localisée par la transition de l'équivalence en AT et GC dans l'ADN. Elle renferme une région intergénique AT-riche de 40 pb qui démontre une forte analogie avec pFQ11.

Mots clés : analyse ORI, RT–PCR, Glimmer, séquence d'ADN.

[Traduit par la Rédaction]

Introduction

Frankia are capable of fixing atmospheric dinitrogen in either the free-living state or in symbiosis with host plants representing over 200 species distributed among 24 genera and 8 families (Nazaret et al. 1991). Actinorhizal host plants are trees and woody shrubs that are distributed worldwide but occur predominantly in temperate areas, higher elevations, and the tropics (Tjepkema et al. 1986). These plants are generally pioneer species colonizing ecologically disturbed or nitrogen-poor sites. The annual nitrogen fixation rates for actinorhizal plants range from 2 to 362 kg of nitrogen per hectare (Stowers 1987). They therefore contribute significant amounts of reduced nitrogen to the ecosystems that they occupy.

Frankia can be quite promiscuous with respect to host-plant selection. Single strains of *Frankia* can nodulate plants from different families. A coevolution of the host plants and *Frankia* is suggested by molecular phylogenetic analysis (Jeong et al. 1999). Their broad host range suggests the biochemical demands on the symbiotic partner may be lower than those of *Rhizobia.* This hypothesis is supported by numerous observations that indicate the nodulation process is genetically and biochemically distinct from the more well studied *Rhizobia* (Guan et al. 1998). Therefore, a better understanding of *Frankia* molecular biology may provide insights into alternative host recognition and (or) colonization mechanisms of nodulation. Biological nitrogen fixation is the most likely avenue to reduce our current demand for fossil fuels in the production of nitrogen fertilizers. *Frankia*

Received November 10, 2000. Revision received February 22, 2001. Accepted February 28, 2001. Published on the NRC Research Press Web site at http://cjm.nrc.ca on June 28, 2001.

T.R. John, J.M. Rice, and J.D. Johnson.[1] Department of Molecular Biology, University of Wyoming, Laramie, WY 82071, U.S.A.

[1]Corresponding author (e-mail: jdjohn@uwyo.edu).

 DOI: 10.1139/cjm-47-7-608

represent an untapped reserve of potential genetic resources in this area.

Compared with *Rhizobia*, little is known of *Frankia* molecular biology and genetics (Benson and Silvester 1993; Mullin 1996). Only a few *Frankia* genes have been cloned and no mutants have been identified. Further, no tools such as transformation, transduction, or conjugation are available for genetic manipulation of these organisms. The inability to do genetics is a major block to progress in understanding *Frankia* at all levels of investigation. The availability of a mechanism to manipulate *Frankia* genetics would enable the advancement of many studies of these interesting microbes. To facilitate development of a transformation system, we have cloned and sequenced a plasmid of 22.4 kb that is present in some but not all members of this genus (Simonet et al. 1985). An 8.5-kb plasmid, typically found in the same isolates, has already been sequenced (Kong et al. 1997), thereby affording the opportunity to search for similarities that can provide information about replication functions. Identification of sequences that function as origins of replication (ORI) for *Frankia* plasmids was one goal of the project. Vectorial analysis of intrastrand AT and GC equivalences must be used to localize ORI regions because no functional test is currently available (Benson and Silvester 1993; Mullin 1996). A second goal was to search for genes that are associated with plasmid stability and replication. Such functions can provide important clues about how to manipulate these plasmids to create cloning vectors for *Frankia*.

Materials and methods

Cell growth and plasmid isolation

Frankia strain CpI1, obtained from Dr. John Torrey (Harvard University), was grown in a stationary culture in M6B+ media (Fontaine et al. 1986) for 2–4 weeks. Plasmid DNAs were isolated by CsCl centrifugation (Normand et al. 1983).

Cloning and DNA sequence analysis

The larger plasmid, isolated from CpI1, was cleaved by *Bam*H1 into fragments of 5.2, 4.8(2), 2.9, 2.8, and 1.8 kb that were resolved by agarose gel electrophoresis (Simonet et al. 1985). These fragments, along with an additional 0.2-kb fragment not apparent on agarose gels, were cloned in the vector pUC13 using *Escherichia coli* JM83 as the host. Plasmid DNAs for use as sequencing templates were purified with Wizard Mini-prep kits (Promega, Madison Wis.). Reactions were done as recommended for the d-Rhodamine cycle-sequencing kits (Perkin Elmer, Foster City, Calif.) and were analyzed on an ABI model 377 automated DNA sequencer. Restriction fragments were ordered using polymerase chain reaction (PCR) amplification across the *Bam*H1 cleavage sites, followed by direct sequencing of the PCR products. Regions within the 5.7-, 5.3-, and 3.0-kb clones that were refractory to direct sequence analysis were amplified by PCR using Vent DNA polymerase (NEB, Beverly, Mass.), cloned in pBSIISK$^+$ (Stratagene, La Jolla, Calif.), and sequenced. In the finished sequence, these regions are positions 2681–3207, 7452–8149, and 15 682 – 16 159. One of the 4.8-kb fragments could not be recovered as a clone in *E. coli*. PCR amplification using primers in the adjacent fragments was not successful with either Taq or Vent polymerases without the addition of PCR$_X$Enhancer (Life Technologies, Grand Island, N.Y.). Reactions including this product gave rise to a 5.0-kb product. This fragment could not be cloned in pBSIISK$^+$, pTZ18R (Pharmacia, Piscataway, N.J.), or pCR-TOPO (Invitrogen, Carlsbad, Calif.). The PCR fragment was cleaved with *Bam*H1+*Sph*1, *Bam*H1+*Taq*1, *Taq*1, or *Xma*1, and the digestion products were cloned into pBSIISK$^+$ or pTZ18R and were sequenced. All regions were sequenced from at least two independent clones and have ≥2 passes over each base.

The annotated sequence is available at GenBank accession No. AY027524.

Open reading frame identification

The locations of open reading frames (ORFs) encoding ≥100 amino acids were first determined using MacVector software (Oxford Molecular, Oxford, U.K.). These ORFs were sorted based on three criteria: (*i*) effective codon use (Wright 1990), (*ii*) G+C abundance in the 3rd position of the codons (Sueoka 1988), (*iii*) and the difference in G+C abundance between the 3rd and 2nd positions. The Glimmer algorithm for microbial gene identification (Delcher et al. 1999; Salzberg et al. 1998) was also used to search for coding regions. The program was trained using the *Mycobacterium tuberculosis* genomic sequence (prerelease version from <http://www.tigr.org>; February 8, 2000) because of the comparable G+C content, 65.5%.

RNA extraction

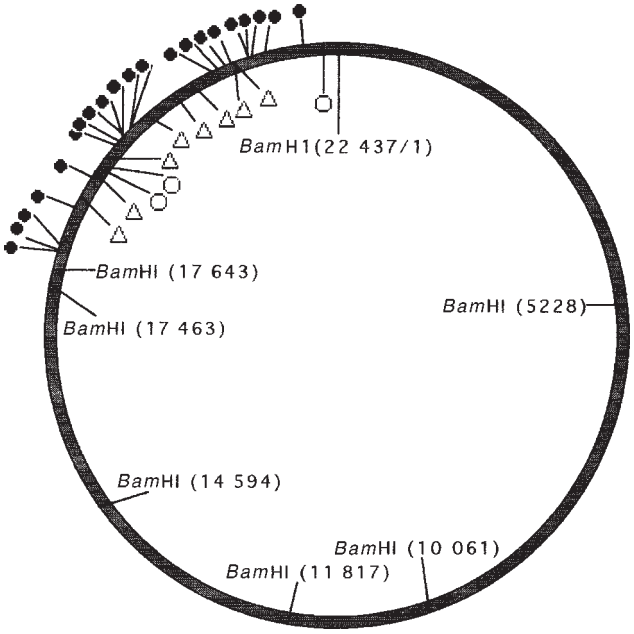
Total RNA was prepared from *Frankia* CpI1 cells grown for 3 weeks at 25°C in media lacking combined nitrogen. Cells from 1L cultures were collected by centrifugation for 10 min at 17 400 × *g* at 4°C. The pellet was washed in 2.3 mL H_2O treated with diethyl pyrocarbonate (DEPC) then pelleted at 69 000 × *g* in a Beckman TLA-100.3 rotor for 15 min at 20°C. Cell yields ranged from 150 to 300 mg/L of culture.

Washed cells were suspended at 0.75–1.0 mg/μL in 3 mg/mL of lysozyme, then 200 μL of 50 μg/mL lysostaphin (Ambion, Austin, Tex.) was added, and the mixture was incubated for 15 min at 37°C to facilitate lysis. Total RNA was extracted using the supplier's basic protocol for RNA isolation with RNAWIZ^{TM}(Ambion). Four millilitres of RNAWIZ^{TM} solution was added to lysed cells and the mixture was vortexed vigorously for 10 s. Particulates were broken up by forcing the mix through an 18-gauge needle, 800 μL of chloroform was added, and the sample was vortexed until homogeneous. Following a 5-min incubation at room temperature, the sample was centrifuged for 15 min at 11 300 × *g* at 4°C, after which the aqueous layer was transferred to a fresh tube, diluted with an equal volume of DEPC-treated H_2O, and RNA was precipitated by adding an equal volume of isopropanol. The precipitate was recovered by centrifugation at 12 700 × *g* for 10 min at 4°C. The pellet was washed by gently vortexing in 70% (v/v) ethanol, recentrifuged for 2 min, then air dried, and dissolved in 70 μL of 0.1 mM EDTA. Finally, the sample was treated with 4U DNaseI (Ambion) and quantitated by measuring absorbance at 260 nm.

Reverse transcriptase–polymerase chain reaction

Transcription of sequences representing pFQ12 ORFs was determined by reverse transcriptase–polymerase chain reactions (RT–PCR). Superscript II reverse transcriptase was used to prepare cDNA from 1 μg of unfractioned *Frankia* RNA with 0.25–1.25 μg of random decamers (Ambion) following the protocol supplied with the enzyme (Life Technologies). Vent DNA polymerase was used for PCR amplification from either a 1/20 aliquot of the RT reaction, 350 ng of total *Frankia* RNA without a RT reaction (negative control), or 20 fmol of a corresponding pFQ12 DNA clone (positive control) as template and 20 pmol each of oligodeoxynucleotide primer pairs that hybridize within pFQ12 ORFs using conditions recommended by the supplier (NEB) with the addition of 5% (v/v) dimethyl sulfoxide (DMSO). Products from RT–PCR reactions with the same electrophoretic mobility as those of the corresponding positive control reaction were excised from the

Fig. 1. Cloning and sequencing of the 22.4-kbp plasmid. *Bam*H1 fragments from the plasmid were cloned in *Esherichia coli* and identified by their size and end sequences. The fragments were ordered by PCR amplification across borders and confirmed by DNA sequence analysis of the amplification products. The *Bam*H1 fragment representing positions 17 643 to 22 437 could not be cloned. A PCR copy of the region was cleaved with *Xma*1(△), *Taq*1(●), or *Sph*1(○) and the resulting fragments were cloned, sequenced, and assembled by overlap.



acrylamide gel and analyzed by DNA sequencing as described previously.

Southern blot

Plasmids carrying cloned fragments of pFQ12 were digested with restriction endonucleases and the fragments were separated by electrophoresis on 1.2% (w/v) agarose, denatured, neutralized, transferred to a nylon membrane by capillary, rinsed in 2× SSC (1×: 0.15 M sodium chloride plus 0.015M sodium citrate), and baked in vacuo for 1 h at 80°C.

A ^{32}P-labelled cDNA probe was prepared from 5 µg of total *Frankia* RNA by random-primed synthesis essentially as described by (Richmond et al. 1999).

Size markers were prepared from *Hin*dIII-digested phage λ DNA via fill-in reactions using [α-^{32}P]dCTP and unlabelled dATP, dGTP, and dTTP with Klenow DNA polymerase following the enzyme supplier's protocol (NEB). The ^{32}P-labelled DNA was separated from unincorporated [^{32}P]dCTP using a NENsorb column.

Filters were probed with a mixture containing 6×10^8 dpm *Frankia* probe, 2.5×10^5 dpm λ DNA marker, and 20 µg of unlabelled *E. coli* rRNA (a generous gift from Dr. S. Franklin) essentially as described (Bock et al. 2001).

ORI identification

The locations of replication origins in both pFQ12 and pFQ11 were determined using vectorial analysis of AT and GC equivalences (Lobry 1996). The original algorithm was modified to provide sequence positions to identify inflection points (T. Phang, University of Wyoming, personal communication).

Results

Cloning and DNA sequencing

Plasmid DNAs from *Frankia* strain CpI1 were purified by equilibrium centrifugation in CsCl gradients and then digested with *Bam*H1. This enzyme does not cleave pFQ11 (Kong et al. 1997) but produces six fragments from pFQ12, which are visible following agarose gel electrophoresis (Simonet et al. 1985). The *Bam*H1 digestion fragments were cloned and end sequenced to identify restriction fragments representing all size classes derived from pFQ12: 5.2, 4.8(2), 2.9, 2.8, and 1.8 kb. Clones carrying *Bam*H1 fragments of 0.2–0.5 kb were also recovered and sequenced. Six independent clones of the 4.8-kb size class were screened and found to be identical. No clone representing the second 4.8-kb fragment was ever identified. A representative clone for each of the other fragments was sequenced by primer walking. The *Bam*H1 fragments were ordered using PCR with intact pFQ12 as a template and primers were directed toward each of the *Bam*H1 cleavage sites. Each PCR product was sequenced to verify contiguity of the restriction fragments. This assembled sequence represents positions 1 – 17 463 (Fig. 1).

A 0.2-kb *Bam*H1 fragment not visible on agarose gels but isolated from the clone bank was positioned by PCR and represents nucleotides 17 464 – 17 643 in the finished sequence.

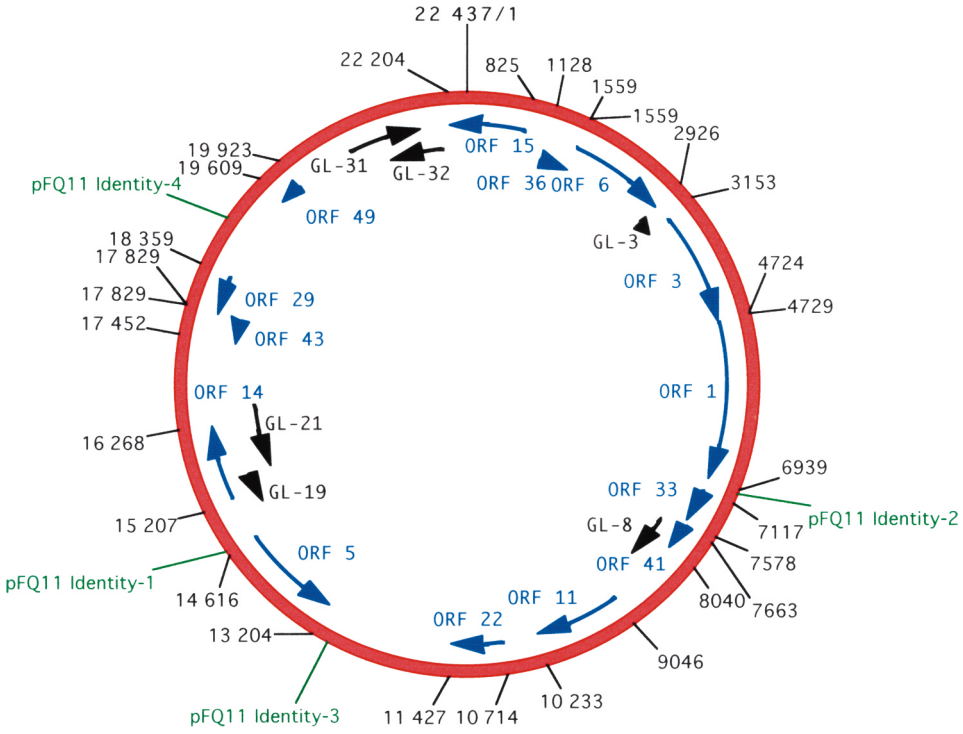
Sequence analysis of the remaining DNA was complicated because the representative *Bam*H1 fragment could not be recovered as an *E. coli* clone. Amplification of the region by PCR using intact plasmid as template and primers complementary to flanking sequences gave rise to a 5.0-kb product. Repeated attempts to clone this fragment failed. Aliquots of the 5.0-kb fragment were then digested with *Taq*1, *Xma*1, or *Bam*H1 plus *Sph*1. Locations of these cleavage sites are identified in Fig. 1. The resulting fragments were cloned and used as templates for DNA sequence analysis to generate the finished sequence from positions 17 644 – 22 437. Continuity between positions 1 and 22 437 was verified by PCR amplification and DNA sequencing across the junction.

The plasmid has an unusually high G+C content, 76%, when compared with five *Streptomyces* plasmids whose complete sequences are available in GenBank, 69%–73% G+C, or the *Frankia* plasmid pFQ11, 72% G+C. Many sequencing reactions were inhibited, apparently by encountering clusters of G and C that could give rise to very stable intrastrand base-paired regions. Numerous direct and inverted repeats were also identified (Fig. 7).

Gene identification

The 14 ORFs that met the original criteria defined in Materials and methods are numbered largest to smallest and are positioned on a circular map of the plasmid (Fig. 2). Additional coding regions, identified by the Glimmer algorithm, are also positioned and labelled with the prefix GL. There was substantial, but not complete, agreement between the two methods of analysis. In addition to recognizing some completely new ORFs, the Glimmer program also extended the start site of some ORFs identified by the MacVector software (Fig. 2). A potentially very large, virtually contiguous coding region exists from positions 1128 through to 6939,

Fig. 2. ORF map of pFQ12. The DNA sequence was analyzed using MacVector and Glimmer software to identify potential protein coding regions in the plasmid. The ORFs identified by computer analysis were further screened based on size, G+C distribution, and codon use bias to determine which were most likely to be bona fide genes. Genes uniquely identified by the Glimmer system are identified by the prefix GL. The location of four short regions of pFQ12 with DNA sequence identity to segments of pFQ11 are also noted.



including five different genes. The region from positions 825 through to 17 452 or 15 207 appears to be transcribed primarily in a counterclockwise direction, as presented. This suggests that the region 825–1128 harbors promoters that direct transcription in both directions around the circle. A second such region appears to exist between positions 14 616 and 15 207.

To determine whether any of the ORFs were bona fide genes, RNA was isolated from *Frankia* cells carrying pFQ12. Attempts to do Northern blots with RNA prepared by a variety of methods all failed. Electrophoretic analysis of RNA preparations on denaturing, agarose gels suggested they were partially degraded.

To circumvent this problem, two approaches were used. *Frankia* RNA was used in reverse transcriptase reactions with random decamers as primers. The resulting cDNA was then amplified by PCR using primer pairs specific for individual ORFs (Fig. 3A). When PCR products of the expected size were produced from these reactions, that DNA was used as a template for DNA sequencing to determine whether the PCR product represented the pFQ12 ORF. This approach confirmed that six of the ORFs, 1, 3, 15, 22, GL-31, and GL-32, were transcribed into RNA and presumably encode proteins (Figs. 2 and 3A). To complement and extend the RT–PCR results, clones representing 14 of the ORFs were digested with restriction endonucleases to release internal fragments of the coding regions. The fragments were separated by agarose gel electrophoresis and blotted onto a charged nylon membrane. The membrane was probed with ^{32}P-labelled cDNA prepared by reverse transcriptase copying of the *Frankia* RNA in the presence of [α^{-32}P]dCTP. A photograph of the agarose gel prior to blotting and a fluorogram of the filter after probing and washing are presented (Figs. 3B and 3C). DNA bands that uniquely represent a single ORF are identified by numbers corresponding to Fig. 2. In some lanes, other bands include mixtures of additional ORF sequences. The hybridization results confirm the transcription of ORFs 1, 3, 15, 22, and GL-31 and further provide support for the expression of ORFs 5, 14, and 29 (Figs. 2 and 3C). The DNA fragments representing ORF 14 in the Southern blot also include sequence from the overlapping ORFs GL-19 and GL-21 (Fig. 2). Hybridization to these fragments may be due to transcription of any or all of these three ORFs. A summary of the mRNA analyses is presented in Table 1.

Searches of the GenBank database using the BLASTP algorithm with putative translation products of the ORFs that were expressed as RNA, in some cases identified very similar sequences in other bacteria (Table 2). A membrane protein known to be involved in λ phage tail fiber assembly strongly resembles the ORF 5 product. A protein encoded by ORF 3 is very similar to the TrbL transfer factor from plasmid RK2. The product of ORF 15 is a membrane protein related to a product of the herpes virus genome. ORF 22 encodes a protein similar to a gene found in *Streptomyces coelicolor* and the ORF 29 protein has some similarity to a *Sinorhizobium* mobilization protein.

A close resemblance, BLASTP P(N) = 7.4×10^{-14}, exists between the gene products of GL-32 in pFQ12 (Fig. 2) and ORF 2 in pFQ11 (Kong et al. 1997). These two *Frankia* pro-

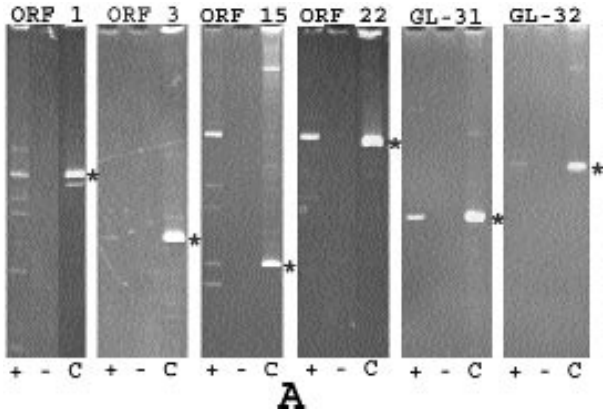


Fig. 3. Detection of RNA transcribed from pFQ12. Total RNA was isolated from *Frankia* CpI1 cells as described in Materials and methods. To determine whether any of the ORFs identified in Fig. 2 were expressed as RNA, two approaches were used: (*i*) RT–PCR amplification using primers embedded within ORFs or (*ii*) probing of a Southern blot of pFQ12 fragments with ^{32}P-labelled cDNA. (A) Polyacrylamide gel electrophoresis of PCR products. A reverse transcription of the RNA preparation was done using random primers to create a cDNA library. PCR amplifications using aliquots of the cDNA were then done using primer pairs embedded within individual ORFs (+). Control reactions, using as template either the RNA preparation without reverse transcription (–) or a plasmid DNA containing a fragment representing the ORF (C), were also done. Reaction mixtures were analyzed by polyacrylamide gel electrophoresis. DNA bands in RT–PCR reactions corresponding in size to the expected fragment in a positive control reaction (*) were excised and subjected to DNA sequence analysis. Numbers above each group of three gel lanes correspond to the ORF numbers in Fig. 2. (B) Agarose gel electrophoresis of recombinant plasmids carrying pFQ12 segments. Recombinant plasmids carrying fragments of pFQ12 that correspond to ORFs were digested with restriction endonucleases and the reaction mixtures were separated by agarose gel electrophoresis. Bands corresponding to unique ORFs are identified by the numbering system used in Fig. 2. Phage λ DNA fragments digested with *Hin*dIII are included as size markers. (C) Autoradiogram from Southern blot of pFQ12 DNA fragments. DNA fragments from the gel in panel B were blotted onto a nylon membrane then probed with ^{32}P-labelled cDNA synthesized using reverse transcriptase with the *Frankia* RNA as a template. Numbers represent ORFs as identified in Fig. 2. Phage λ DNA digested with *Hin*dIII and probed with ^{32}P-labelled λ DNA is included to provide size markers.

Table 1. Detection of transcripts from pFQ12.

ORF No.	PCR	BLOT
1	Yes	Yes
3	Yes	Yes
5	No	Yes
6	No	No
11	No	No
14*	No	Yes
15	Yes	Yes
22	Yes	Yes
29	nd	Yes
33	nd	No
36	nd	nd
41	nd	No
43	nd	No
49	No	No
GL-3	No	nd
GL-19	No†	?
GL-21	No†	?
GL-31	Yes	Yes
GL-32	Yes	nd

Note: nd, not determined; ?, ambiguous.
*May be GL-19 or GL-21.
†Primers used did not include ORF 14 sequence.

teins both have about the same degree of resemblance to the *Streptomyces* korSA protein (Hagege et al. 1993) as they do to each other. Proteins encoded by ORFs 1, 14, 36, 41, 43, 49, GL-3, GL-8, GL-21, and GL-31 do not resemble any-

Table 2. GenBank identities with pFQ12 ORFs.

ORF No.	Mol. W.	pI	Principal identities with known proteins	Domains/Sites/Composition
1	78 013	5.04	None	ATP/GTP binding site HTH in AraC transcription regulator
3	52 208	12.6	3.00×10^{-13} TrbL transfer factor from plasmid RK2	Multiple membrane spanning domains
5	48 494	4.87	5.40×10^{-29} λ Tail fiber assembly protein 5.40×10^{-29} *Escherichia coli* putative membrane protein	36% Ala
6	46 788	5.07	1.10×10^{-10} *Streptomyces coelicolor* putative secreted protein 9.80×10^{-10} *Enterobacter* plasmid R751 KfraA protein	ATP/GTP binding site 27% Ala
11	39 913	10.5	7.00×10^{-27} *Mycobacterium tuberculii* hypothetical protein 3.70×10^{-22} *Bacillus subtilis* SpoJ protein 2.30×10^{-20} chromosome partitioning protein	8 AA identical to FBPase active site
14	37 718	12.4	None	No remarkable features
15	36 647	12.9	4.00×10^{-12} herpes virus membrane glycoprotein	23% Ser + Thr; 21% Arg + His
22	24 874	5.7	3.60×10^{-15} *Streptomyces coelicolor* hypothetical protein	7 potential transmembrane domains
29	18 939	10.7	7.40×10^{-2} *Sinorhizobium* mob protein	No remarkable features
33	17 283	13.0	5.80×10^{-6} *Pseudomonas* hypothetical gene	7 Cys and 37 Arg residues
36	14 944	8.8	None	20% Leu + Val
41	12 870	13.0	None	36% Ala + Ser
43	13 185	13.2	None	32% Ala + Gly; 18% Arg
49	10 393	3.6	None	40% Ala + Gly + Val; 12% Glu + Asp
GL-3	7 883	4.6	None	26% Ala
GL-8	24 544	11.5	None	47% Ala + Pro + Arg
GL-19	15 526	7.3	3.6×10^{-3} *Mycobacterium* hypothetical protein	7 potential transmembrane domains 48% Ala + Gly + Leu
GL-21	35 390	11.3	None	35% Ala + Gly
GL-31	35 352	12.5	None	26% Ser + Gly
GL-32	29 804	9.6	2×10^{-14} *Frankia* pFQ11 hypothetical protein 4×10^{-12} *Streptomyces* korSA regulatory protein	11% Pro; 28% acidics + basics

thing currently found in the database. Proteins similar to those encoded by ORFs 6 and 33 are present in the GenBank database, even though neither RT–PCR nor Southern blotting experiments provided evidence for transcription of these regions of pFQ12 (Table 2).

ORI identification

A primary goal of the investigation of pFQ12 was to identify a sequence or region that is capable of functioning as an origin of replication in *Frankia*. The availability of the base sequence of another *Frankia* plasmid, pFQ11 (Kong et al. 1997), afforded an opportunity to do a matrix comparison of two plasmids searching for identities. Restriction endonuclease mapping of the two plasmids suggested they would be quite different (Simonet et al. 1985). This was confirmed by the sequence comparison. Only four small (40–90 bp), dispersed islands of identity with pFQ11 were found. The locations of these are noted in Fig. 2. Alignments of the pFQ12 and pFQ11 cognates are presented in Fig. 4.

An algorithm has been developed to identify bacterial ORI sequences based on changes in intrastrand AT and GC contents (Frank 2000; Lobry 1996). This program was used to localize putative ORI sequences in both plasmids (Fig. 5). The ORI regions are between positions 14 500 and 15 000 in pFQ12 and positions 5800 and 5850 in pFQ11 (Fig. 5). This region of pFQ12 is intergenic and includes the most A+T-rich segment of the plasmid (14 630 – 14 670), which is identity 1 in pFQ11 (Fig. 4). Two 15-bp inverted repeats, 14 632 – 14 646 and 14 655 – 14 669, overlap this region of identity (Figs. 6 and 7). In pFQ11, the ORI region identified by the algorithm is G+C rich, 82%, and located within an ORF identified by the Glimmer program. The region of pFQ11 that is quite similar to the A+T-rich segment of pFQ12, likely to form at least a part of that ORI, is centered around position 6470 (Fig. 4).

Identity 2 is also somewhat A+T rich (Fig. 4). In pFQ12, it is located 120 bp from the 5′ end of ORF 33 (Fig. 2). In pFQ11, the sequence is embedded in ORF 3.

Identity 3 is G+C rich and located in intergenic regions in both plasmids. In both cases, the sequences lie at a position such that flanking ORFs are transcribed in convergent directions.

Identity 4 is modestly A+T rich given the background of both plasmids. Each sequence is located in an intergenic region with no obvious relationship to adjacent ORFs.

Repeated sequences

The abundance of large, repeated sequences in the plasmid and their asymmetric distribution is striking (Fig. 7). There are two large groups of direct repeats. One is a cluster of repeats with lengths of 50, 31, 23, and 22 bp grouped around positions 17 000 – 17 800 (Fig. 7). The other is a collection of tandem repeats of 22–23 bp, grouped around position 20 600, which may be responsible for the difficulty in cloning large fragments that contained the region 17 643 – 22 437.

Fig. 4. DNA sequence identities between pFQ11 and pFQ12. A matrix comparison of the two *Frankia* plasmid DNA sequences revealed only four small regions of similarity. The positions of these on the pFQ12 map are shown in Fig. 2. The aligned sequences and their locations in each plasmid are presented.

IDENTITY #1

```
            6490          6480          6470          6460          6450
pFQ11       CC   ggTCccAAAAcCAGATTTCCGGAAACAGAAAaC  cGgTTTTg
pFQ12       CCTACTCTAAAAACAGATTTCCGGAAACAGAAATCTGTTTTTA
            14 630        14 640        14 650        14 660        14 670
```

IDENTITY #2

```
                  7590          7600          7610          7620
pFQ11       ACGCAAG cC CGAcGACGCAAGGGATGACGCAAgCCACGA
pFQ12       ACGCAAGGCCGATGACGCAAGGGATGACGCAACCCACGA
            6940          6950          6960          6970

                  7630          7640          7650
pFQ11       CGCAAGGCATGACGCAAGG  aGGATGCAAGg
pFQ12       CGCAAGGCATGACGCAAGGCGGATGCAAGC
            6980          6990          7000
```

IDENTITY #3

```
                        5070          5080          5090
pFQ11       GGTCGGCGTACGCCTC  cCTTATGTCGCGGAGTGC  tTGGG gG
pFQ12       GGTCGGCGTACGCCTCTCTTATGTCGCGGAGTGCCTGGGAG
                  12 910        12 920        12 930        12 940

            5100          5110          5120
pFQ11       GTGCCCAGCGG tCCGC cGACCTGC  gCCAGC
PFQ12       GTGCCCAGCGGCCCGCTGACCTGCACCAGC
                  12 950        12 960        12 970
```

IDENTITY #4

```
                  3910          3900          3890          3880
pFQ11       GGTGTGCAGCGGGc TGTGCAACg CCGTGCACGACGGTGTGCAA
pFQ12       GGTGTGCAGCGGGGTGTGCAACCCCGTGCACGACGGTGTGCAA
                  18 960        18 970        18 980        18 990

            3870          3860          3850          3840
pFQ11       ACCGGTGTGCAAg a t CcGTGCATGCACGGCCCCAACGCGC
pFQ12       ACCGGTGTGCAACGACtGTGCATGCACGGCCCCAACGCGC
            19 000        19 010        19 020        19 030
```

There are also 14 pairs of inverted repeats of 15–24 nucleotides each, in pFQ12. These appear to be dispersed throughout the sequence with a notable paucity in the region from 16 to 24 000. A pair of 11-bp inverted repeats with copies starting at positions 1579 and 15 403 are especially interesting because they are each flanked by identical 4-bp direct repeats.

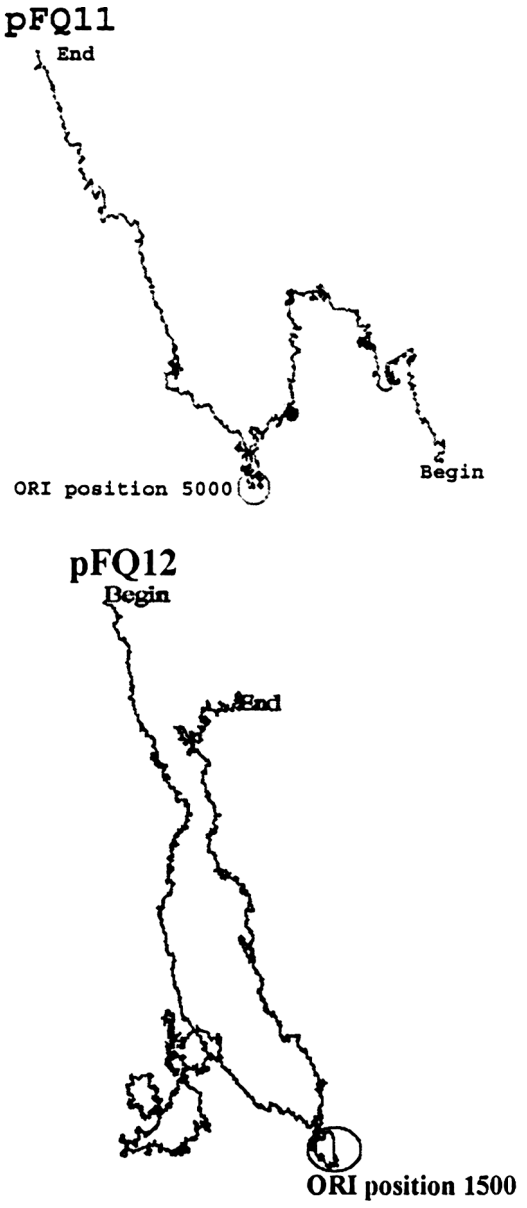
Eight palindromes from 10 to 28 bp are also dispersed around the plasmid. Again, the region between positions 16 400 and 24 000 is free of these structures.

Vectorial analysis identified regions in the pFQ12 sequence that appear to be the result of internal inversions because of the loops that are generated (Fig. 5). A segment covering 8.8 kbp appears to have undergone at least three such events.

Discussion

The *Frankia* plasmids pFQ11 and pFQ12 were first described by Normand et al. (1983) as 7- and 19-kb class molecules present in strain CpI1. Subsequent cloning and restriction endonuclease mapping of the related plasmids from strain ArI3 suggested that the two classes of plasmids were unrelated to each other (Normand et al. 1985) but were quite similar between the two strains (Simonet et al. 1985). The DNA sequence of the two plasmids has now confirmed the divergence but also identified four short regions of identity between the two plasmids at the level of DNA sequence

Fig. 5. Vectorial representation of pFQ12 and pFQ11 sequences for the detection of replication origins. The AT and GC intrastrand equivalence has been shown to switch polarity at the origin of replication in prokaryotic genomes. A vectorial representation of this switch can be used to localize the ORI site. This method was applied to pFQ12 and pFQ11, and clear reversals were apparent with both plasmids. The circled region in each sequence identifies the ORI region of the plasmids.



(Figs. 2 and 4). There is also a strong similarity between the amino acid sequences encoded by one ORF on each plasmid. This identity is not recognizable at the DNA level.

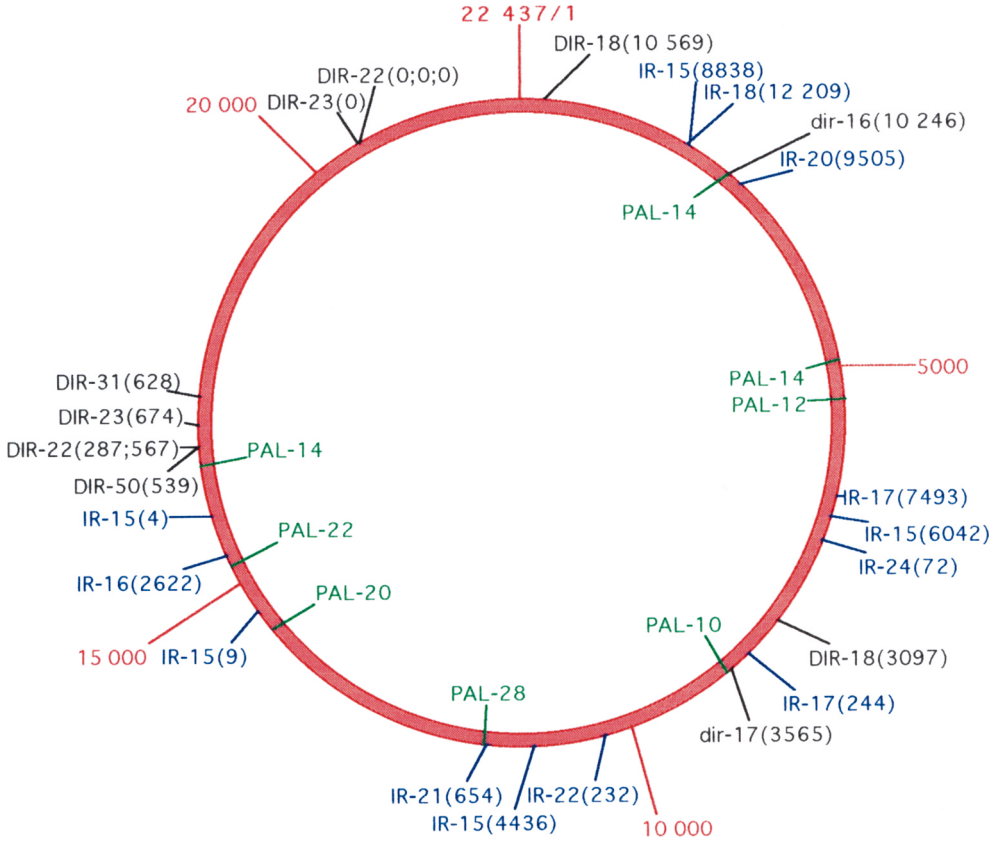
At the protein level, each plasmid has an identity with the korSA protein encoded by the conjugative, integrative plasmid pSAM2 in *Streptomyces ambofaciens.* The korSA protein down regulates the intermycelial transfer of the plasmid (Hagege et al. 1993). The GL-32 and ORF 2 proteins of pFQ12 and pFQ11 may well have a similar function, suggesting that both plasmids may be capable of conjugative transfer. Further, a chromosomal integration site for pSAM2

Fig. 6. Alignment of plasmids pFQ11 and pFQ12 with a fragment of *Frankia* strain ArI3 DNA. Several fragments of DNA cloned from *Frankia* strain ArI3 were shown to have promoter activity in *Escherichia coli* (Cournoyer and Normand 1994). One of these fragments has substantial sequence identity to a segment of pFQ12 DNA that has been noted as also being present in pFQ11 (Fig. 4). The region of identity includes an imperfect inverted repeat spanning about 50 bp that is underlined in the ArI3 sequence. Positions identical in all three sequences are noted with an asterisk.

```
SOURCE     POSITION                    SEQUENCE
pFQ11        6 393     GA--GCCCGCGACGTC-GACGAGCACCAGGACGCCGGCGCCGAGCCCCTCCCCACCC
pFQ12RC    14 720     GC--GCACCGCCGGCC-CACGATCATCCTCAAACCTTCCCGGA---CCCTCCCACCC
ArI3          46      GACAGCGACCTCCGCCAGGAGATCATCCTCAAACCTTCCCGGA---CCCTCCCACCC
                      *   **       * *    ** ** *   *  **  * * **   **  *******

pFQ11        6 447     CTTCAAAACCGGTTTTCTGTTTCCGGAAATCTGGTTTTGGGACCGGACGATGCCC
pFQ12RC    14 670     CT--AAAAACAGATTTCTGTTTCCGGAAATCTGTTTTTAGAGTAGGGTGGGGATC
ArI3         100     CT--AAAAACAGATTTCTGTTTCCGGAAATCTGTTTTTAGAGTAGGGTGGGGATC
                     **  **** * * ******************** **** *    **  *  *  *
```

Fig. 7. Direct and inverted repeats plus palindromes in pFQ12. The type, locations, and size of direct and indirect repeats ≥15 bp in the plasmid are identified as indicated by the following examples: DIR-18 (10 569) indicates a direct repeat of 18 bp with the 2nd copy starting 10 569 bp in the clockwise direction from the first. If there are more than two copies of the sequence, multiple positions are indicated in parentheses; IR-15 (8838) indicates an inverted repeat of 15 bp with the 2nd copy starting 8838 bp in the clockwise direction from the 1st; pallindromes ≥10 bp are identified and indicated by the following convention: PAL-14 indicates a palindrome with a total length of 14 bp at that site.



in *Frankia* DNA has been identified by cloning in *Streptomyces lividans* (Alegre et al. 1994). These observations suggest a very attractive possibility for the construction of vectors capable of introducing and stabilizing exogenous DNA in *Frankia* cells.

Based on sequence identities with other organisms, it is also likely that pFQ12 encodes at least three other proteins, ORFs 3, 6, and 11, involved in plasmid stability and transmission. A more modest similarity to a mobilization protein from *Sinorhizobium* is seen for the ORF 29 product. All of these ORFs, except number 6, are expressed in vivo (Fig. 3). The homologue of ORF 6, KfrA, is thought to function as an autogenous repressor (Jagura-Burdzy 1992). The presence of these protein functions suggests that pFQ12 may be transmissible to other cells via conjugation, a very useful property for a potential cloning vector.

Organization and expression of genes in *Frankia* is a largely unexplored area. The genomes and plasmids in these bacteria are very G+C rich, leading to strong codon bias (Harriott et al. 1995) as well as an abundance of amino acids represented by G+C-rich codons (Table 2). The identification and characterization of *Frankia* genes involved in host

recognition and nodulation will likely require thorough characterization of promoter sequences and development of cloning vectors specifically adapted for this organism. The promoter regions associated with the genes identified by transcriptional analysis of pFQ12 (Fig. 3) are good candidates for creating fusions to express reporter genes, such as green fluorescent protein.

The development of cloning vectors will likely also involve identification of sequences capable of functioning as ORI regions. A candidate for this function is an A+T-rich region of identity between the two plasmids that both map near ORI regions predicted by in silico analysis (Fig. 5). Further, a sequence with substantial similarity, but not complete identity, to these putative ORI regions (Fig. 6) was cloned from *Frankia* strain ArI3 and demonstrated to have promoter activity in *E. coli* (Cournoyer and Normand 1994). Strain ArI3 carries plasmids identical, at the level of restriction endonuclease mapping, with pFQ11 and pFQ12 (Simonet et al. 1985). This sequence in pFQ12 is located near the 5′ end of ORF 5 (Fig. 2). The pFQ12 sequence has an imperfect inverted repeat of about 40 bp that includes possible promoter elements (Cournoyer and Normand 1994). The region could therefore also serve as a promoter for ORF 14 that is divergently transcribed with respect to ORF 5. However, this promoter would be about 500 bp upstream from the translational start site. The sequence in pFQ11 is in an intergenic region approximately 2 kbp from the 5′ end of the nearest ORF judged likely to be a protein encoding region by Kong et al. (1997) or the Glimmer algorithm (Delcher et al. 1999).

It is quite possible that the identity region in pFQ11, pFQ12, and ArI3 (Fig. 6) has both ORI and promoter activities. This is known to be the case for OriC promoters in many bacteria, including the closely related *Streptomyces* (Zakrazewska-Czerwinska and Schrempf 1995). All known *Streptomyces* plasmids, including pSAM2, that encode a transfer protein in common with both pFQ11 and pFQ12, replicate via a rolling circle mechanism (Suzucki et al. 1997). However, a conserved nicking site in their ORI regions, CTTGGGA or CTTGATA, is not present in the sequences of either pFQ11 or pFQ12.

Although the DNA sequences of the two *Frankia* plasmids are quite different, it is remarkable that four very small but strong identities exist almost exclusively in noncoding regions of both plasmids. This implies that in addition to the ORI function suggested for region 1, the others may also have some function that is not evident from either their sequence or locations.

Analysis of repeated sequences in pFQ12 reveals an unusually high density of these structural elements. Comparison of pFQ12 with the five *Streptomyces* plasmids that have complete sequences in GenBank, indicates that the density of repeats of ≥15 bp in the *Frankia* plasmid is substantially greater, 7.58/1000 bp, than that in *Streptomyces*, 2.76 ± 1.9 / 1000 bp.

A pair of 11-bp inverted repeats each flanked by identical 4-bp direct repeats may mark the boundaries of a transposition event that would have created two-thirds of the plasmid (Fig. 6). A segment bounded by positions 4000 and 12 800 seems to have undergone multiple inversions based on reversals of AT and GC equivalences (Fig. 5) (Lobry 1996). These two observations shed light on the formation and evolution of the plasmid.

A collection of tandem repeats of 22–23 bp grouped around position 20 600 may be responsible for the instability of clones containing the region 17 643 – 22 437. Subclones that separated these regions were stable in *E. coli*.

Acknowledgements

This work was supported in part by a National Science Foundation grant, MCB-9724800, and the Agricultural Experiment Station of the University of Wyoming, College of Agriculture. The intellectual and material contributions of Dr. J.R. Lobry, Ms. Katja Manninen, Mr. Bruce Peterson, Dr. Tzulip Phang, and Mr. Genaro Scavello were very important to the completion of this work.

References

Alegre, M.-T., Cournoyer, B., Mesas, J.-M., Guerineau, M., Normand, P., and J.-L. Pernodet. 1994. Cloning of *Frankia* species putative $tRNA^{Pro}$ genes and their efficacy for pSAM2 site-specific integration in *Streptomyces lividans*. Appl. Environ. Microbiol. **60**: 4279–4283.

Benson, D.R., and Silvester, W.B. 1993. Biology of *Frankia* strains, actinomycete symbionts of actinorhizal plants. Microbiol. Rev. **57**: 293–319.

Bock, J.V., Battershell, T., Wiggington, J., John, T.R., and Johnson, J.D. 2001. Identification of *Frankia* sequences exhibiting RNA polymerase promoter activity. Microbiology, **147**: 499–506.

Cournoyer, B., and Normand, P. 1994. Gene expression in *Frankia*: characterization of promoters. Microbios, **78**: 229–236.

Delcher, A.L., Harmon, D., Kasif, S., White, O., and Salzberg, S.L. 1999. Improved microbial gene identification with Glimmer. Nucleic Acids Res. **27**: 4636–4641.

Fontaine, M.S., Young, P.H., and Torrey, J.G. 1986. Effects of long-term preservation of *Frankia* strains on infectivity, effectivity, and in vitro nitrogenase activity. Appl. Environ. Microbiol. **51**: 694–698.

Frank, A.C., Jr. 2000. Oriloc: prediction of replication boundaries in unannotated bacterial chromosomes. Bioinformatics (Oxford), **16**: 560–561.

Guan, C., Pawlowski, K., and Bisseling, T. 1998. Interaction between *Frankia* and actinorhizal plants. Subcell. Biochem. **29**: 165–189.

Hagege, J., Pernodet, J.-L., Sezonov, G., Gerbaud, C., Friedmann, A., and Guerineau, M. 1993. Transfer functions of the conjugative integrating element pSAM2 from *Streptomyces ambofaciens*: characterization of a *kil-kor* system associated with transfer. J. Bacteriol. **175**: 5529–5538.

Harriott, O.T., Hosted, T.J., and Benson, D.R. 1995. Sequences of *nif*X, *nif*W, *nif*Z, *nif*B and two ORFs in the *Frankia* nitrogen fixation gene cluster. Gene, **161**: 63–67.

Jagura-Burdzy, G., and Thomas, C.M. 1992. KfrA gene of broad host range plasmid RK2 encodes a novel DNA-binding protein. J. Mol. Biol. **225**: 651–660.

Jeong, S.C., Ritchie, N.J., and Myrold, D.D. 1999. Molecular phylogenies of plants and *Frankia* support multiple origins of actinorhizal symbioses. Mol. Phylogenet. Evol. **13**: 493–503.

Kong, R., Xu, X., and Wolk, C.P. 1997. *Frankia* sp. plasmid pFQ11 regulatory protein, transmembrane protein, and partition protein genes, complete cds.: NCBI: GenBank accession No. AF014839).

Lobry, J.R. 1996. A simple vectorial representation of DNA sequences for the detection of replication origins in bacteria. Biochimie**, 78**: 323–326.

Mullin, B., and Benson, D. 1996. Molecular analysis of actinorhizal symbiotic systems: progress to date. Plant Soil, **186**: 9–20.

Nazaret, S., Cournoyer, B., Normand, P., and Simonet, P. 1991. Phylogenetic relationships among *Frankia* genomic species determined by use of amplified 16S rDNA sequences. J. Bacteriol. **173**: 4072–4078.

Normand, P., Simonet, P., Butour, J.L., Rosenberg, C., Moiroud, A., and Lalonde, M. 1983. Plasmids in *Frankia* sp. J. Bacteriol. **155**: 32–35.

Normand, P., Downie, J.A., Johnston, A.W.B., Kieser, T., and Lalonde, M. 1985. Cloning of multicopy plasmids from the actinorhizal nitrogen-fixing bacterium *Frankia* sp., and determination of its restriction map. Gene, **34**: 367–370.

Richmond, C.S., Glasner, J.D., Mau, R., Jin, H., and Blattner, F.R. 1999. Genome-wide expression profiling in *Escherichia coli* K-12. Nucleic Acids Res. **27**: 3821–3835.

Salzberg, S., Delcher, A., Kasif, S., and White, O. 1998. Microbial gene identification using interpolated Markov models. Nucleic Acids Res. **26**: 544–548.

Simonet, P., Normand, P., Moiroud, A., and Lalonde, M. 1985. Restriction enzyme digestion patterns of *Frankia* plasmids. Plant Soil, **87**: 49–60.

Stowers, M.D. 1987. Collection, isolation, cultivation, and maintenance of *Frankia*. *In* Symbiotic nitrogen fixation techniques. *Edited by* G.H. Elkan. Marcel Dekker Inc., New York. pp. 29–53.

Sueoka, N. 1988. Directional mutation pressure and neutral molecular evolution. Proc. Natl Acad. Sci. U.S.A. **85**: 2653–2657.

Suzucki, I., Seki, T., and Yoshida, T. 1997. Nucleotide sequence of a nicking site of the *Streptomyces* plasmid pSN22 replicating by the rolling circle mechanism. FEMS Microbiol. Lett. **150**: 283–288.

Tjepkema, J.D., Schwintzer, C.R., and Benson, D.R. 1986. Physiology of actinorhizal nodules. Annu. Rev. Plant Physiol. **37**: 209–232.

Wright, F. 1990. The effective number of codons used in a gene. Gene, **87**: 23–29.

Zakrazewska-Czerwinska, J.J.M., and Schrempf, H. 1995. Minimal requirements of the *Streptomyces lividans* 66 oriC region and its transcriptional and translational activities. J. Bacteriol. **177**: 4765–4771.